Jeroen van den Hoven
Pieter E. Vermaas
Ibo van de Poel
*Editors*

# Handbook of Ethics, Values, and Technological Design

## Sources, Theory, Values and Application Domains

Springer Reference

# Handbook of Ethics, Values, and Technological Design

Jeroen van den Hoven • Pieter E. Vermaas
Ibo van de Poel
Editors

# Handbook of Ethics, Values, and Technological Design

## Sources, Theory, Values and Application Domains

With 89 Figures and 21 Tables

Springer Reference

*Editors*
Jeroen van den Hoven
Department of Ethics and Philosophy of Technology
Delft University of Technology
Delft, The Netherlands

Pieter E. Vermaas
Department of Ethics and Philosophy of Technology
Delft University of Technology
Delft, The Netherlands

Ibo van de Poel
Department of Ethics and Philosophy of Technology
Delft University of Technology
Delft, The Netherlands

ISBN 978-94-007-6969-4 ISBN 978-94-007-6970-0 (eBook)
ISBN 978-94-007-6971-7 (print and electronic bundle)
DOI 10.1007/978-94-007-6970-0

Library of Congress Control Number: 2015934439

Springer Dordrecht Heidelberg New York London


Printed on acid-free paper

Springer Science+Business Media B.V. Dordrecht is part of Springer Science+Business Media (www.springer.com)

# About the Editors

**Jeroen van den Hoven**
Department of Ethics and Philosophy of Technology
Delft University of Technology
Delft
The Netherlands

**Jeroen van den Hoven** is professor of Moral Philosophy and Vice Dean of the Faculty of *Technology, Policy and Management* at Delft University of Technology. He is Editor in Chief of *Ethics and Information Technology* (Springer). He has published numerous articles on Ethics and ICT. Van den Hoven has received several grants from the Netherlands Organisation for Scientific Research (NWO). He has been advisor to the Dutch Government and European Commission in various roles. Van den Hoven chairs The Responsible Innovation Program of the Netherlands Organisation for Scientific Research and advised the European Commission on the integration of Responsible Innovation into the Horizon 2020 Framework. In 2009, he won the World Technology Award in Ethics.

**Pieter E. Vermaas**
Department of Ethics and Philosophy of Technology
Delft University of Technology
Delft
The Netherlands

**Pieter Vermaas** is associate professor at the Ethics and Philosophy Department of Delft University of Technology. His current research in philosophy of technology focuses on design methods for understanding what design is and for determining how to validate design methods in the traditional domains of engineering, product development, and architecture and in new domains such as business, policy, and the social realm. This research builds on earlier analytic projects on engineering and on the concepts of technical artifact and technical function. These projects resulted in an action-theoretical analysis of the design and use of artifacts and in a philosophical account of technical functions called the ICE theory (*Technical Functions*, Springer, 2010) as well as in a textbook on philosophy of technology (*A Philosophy of Technology*, Morgan and Claypool, 2011). Vermaas is Editor in Chief of the Springer book series *Philosophy of Engineering and Technology*.

**Ibo van de Poel**
Department of Ethics and Philosophy of Technology
Delft University of Technology
Delft
The Netherlands

**Ibo van de Poel** is Anthonie van Leeuwenhoek Professor in Ethics and Technology at Delft University of Technology. Since 1997, he has lectured in ethics and technology for several engineering course programs at Delft University of Technology. His research focuses on responsible innovation, engineering ethics, the moral acceptability of technological risks, values and engineering design, moral responsibility in research networks, and ethics of new emerging technologies like nanotechnology. He is coeditor of the *Handbook of Philosophy of Technology and the Engineering Sciences* (Elsevier, 2009), *Philosophy and Engineering* (Springer, 2010), and *Moral Responsibility: Beyond Free Will and Determinism* (Springer, 2011) and coauthor of *Ethics, Engineering and Technology* (Wiley-Blackwell, 2011). He is also a coeditor of the Springer book series *Philosophy of Engineering and Technology*. Recently, he has received a prestigious VICI grant for his research proposal *New Technologies as Social Experiments: Conditions for Morally Responsible Experimentation* from the Netherlands Organisation for Scientific Research (NWO).

# Contents

# Contributors

**Huib Aldewereld** Delft University of Technology, Delft, The Netherlands

**Frances Brazier** Delft University of Technology, Delft, The Netherlands

**Philip Brey** Universiteit Twente, Enschede, The Netherlands

**Brigitte Burgemeestre** Delft University of Technology, Delft, The Netherlands

Pandar, Amsterdam, The Netherlands

**Aad Correljé** Delft University of Technology, Delft, The Netherlands

**Janet Davis** Computer Science, Grinnell College, Grinnell, IA, USA

**Marc de Vries** TU Delft, Delft, The Netherlands

**Francien Dechesne** Delft University of Technology, Delft, The Netherlands

**Pieter M. A. Desmet** TU Delft, Delft, The Netherlands

**Virginia Dignum** Delft University of Technology, Delft, The Netherlands

**Neelke Doorn** Department of Technology, Policy and Management, TU Delft, Delft, The Netherlands

**Claudia Eckert** Engineering and Innovation, The Open University, Milton Keyes, UK

**Jessica Nihlén Fahlquist** TU Delft / 3TU, Centre for Ethics and Technology, Delft, The Netherlands

**Maarten Franssen** Section of Philosophy, Fac. TPM, Delft University of Technology, Jaffalaan 5, Delft, The Netherlands

**John Grin** Department of Political Sciences, University of Amsterdam, Amsterdam, The Netherlands

**John Groenewegen** Delft University of Technology, Delft, The Netherlands

**Armin Grunwald** Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

**Sven Ove Hansson** Division of Philosophy, Royal Institute of Technology, Stockholm, Sweden

**Paul Hekkert** Department of Industrial Design, Design Aesthetics, TU Delft, Delft, The Netherlands

**Alina Huldtgren** Fachhochschule Düsseldorf, Düsseldorf, Germany

**Joris Hulstijn** Delft University of Technology, Delft, The Netherlands

**Urjan Jacobs** TU Delft, Delft, The Netherlands

**Simeon Keates** School of Engineering, University of Greenwich, London, UK

**Jan Leen Kloosterman** Department of Radiation Science and Technology (RST), Faculty of Applied Sciences, TU Delft, Delft, The Netherlands

**Peter Kroes** Delft University of Technology, Delft, The Netherlands

**Otto Kroesen** TU Delft, Delft, The Netherlands

**Rolf Künneke** Delft University of Technology, Delft, The Netherlands

**Noëmi Manders-Huits** Philosophy Department, TU Delft, Delft, The Netherlands

**Seumas Miller** Australian National University, Canberra, Australia

**Christina Mörtberg** Department of Informatics, Linnaeus University, Växjö, Sweden

**Lisa P. Nathan** University of British Columbia, Vancouver, BC, Canada

**Caroline Nevejan** Delft University of Technology, Delft, The Netherlands

**Philip J. Nickel** Eindhoven University of Technology, Eindhoven, The Netherlands

**Ilse Oosterlaken** Department of Values and Technology, Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands

**Sjef Orbons** Nederlandse Defensie Academie, The Hague, The Netherlands

**Auke Pols** School of Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands

**Wim Ravesteijn** TU Delft, Delft, The Netherlands

**Rob Reuzel** Department for Health Evidence (133), Radboud University Medical Center, Nijmegen, The Netherlands

**Sabine Roeser** TU Delft, Delft, The Netherlands

**Lambèr Royakkers** Technical University of Eindhoven, Eindhoven, The Netherlands

**Daniel Scholten** Delft University of Technology, Delft, The Netherlands

**Lara Schrijver** University of Antwerp, Antwerp, Belgium

**Andreas Spahn** School of Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands

**A. Spahn** School of Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands

**Behnam Taebi** Department of Philosophy, Faculty of Technology, Policy and Management, TU Delft, Delft, The Netherlands

Belfer Center for Science and International Affairs, John F. Kennedy School of Government, Harvard University, Cambridge, MA, USA

**Yao-hua Tan** Delft University of Technology, Delft, The Netherlands

**Nynke Tromp** Department of Industrial Design, Design Aesthetics, TU Delft, Delft, The Netherlands

**Ibo van de Poel** Department of Ethics and Philosophy of Technology, Delft University of Technology, Delft, The Netherlands

**Henk van den Belt** Wageningen University, Wageningen, The Netherlands

**Jeroen van den Hoven** Department of Ethics and Philosophy of Technology, Delft University of Technology, Delft, The Netherlands

**Maja van der Velden** Department of Informatics, University of Oslo, Oslo, Norway

**Gert Jan van der Wilt** Department for Health Evidence (133), Radboud University Medical Center, Nijmegen, The Netherlands

**Pieter E. Vermaas** Department of Ethics and Philosophy of Technology, Delft University of Technology, Delft, The Netherlands

**Joost Vogtländer** TU Delft, Delft, The Netherlands

**Martijn Warnier** Delft University of Technology, Delft, The Netherlands

**Renee Wever** TU Delft, Delft, The Netherlands

University of Limerick, Limerick, Ireland

**Karen Yeung** The Centre for Technology, Ethics, Law and Society (TELOS), The Dickson Poon School of Law, King's College London, London, UK

**Sjoerd D. Zwart** TU Eindhoven, Eindhoven, The Netherlands

# Design for Values: An Introduction

Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel

## Contents



## Design for Values: The Possibility

The design of new products, public utilities, and the built environment is traditionally seen as a process in which the moral values of users and society hardly play a role. The traditional view is that design is a technical and value-neutral task of developing artifacts that meet functional requirements formulated by clients and users. These clients and users may have their own moral and societal agendas, yet for engineers, these are just externalities to the design process. An entrenched view on architecture is that "star" architects and designers somehow manage to realize their aesthetic and social goals in their design, thus imposing their values rather than allowing users and society to obtain buildings and artifacts that meet user and societal values.

This handbook makes available work in ethics of technology and design research that provides an altogether different and more constructive perspective on the possibility to develop technology in accordance with the moral values of users and society at large. In spite of the traditional views, efforts have been made to bring technologies more in sync with our values, and in the handbook these are described. This can be seen as the result of the fact that the stakes of our design

J. van den Hoven (✉) • P.E. Vermaas • I. van de Poel
Department of Ethics and Philosophy of Technology, Delft University of Technology, Delft, The Netherlands
e-mail: m.j.vandenhoven@tudelft.nl; P.E.Vermaas@tudelft.nl; I.R.vandePoel@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_40

endeavors in multiple sectors are so much higher now and design and technology play such an important role in our lives. Moreover, designers themselves are increasingly shifting their perspective toward one in which moral, social, and personal values are to be included in the requirements and in which designers develop products, utilities, and buildings that realize these values. In parallel to this shift, ethicists have articulated the vision that it is important to arrive at morally and societally responsible technological innovations by incorporating moral and societal values into the design processes. It is this position that design for moral and societal values is a possibility that is a defining feature of this handbook. It presents the historical sources of the possibility of closely linking design and values and continues with numerous contributions in which Design for Values is introduced theoretically, explored for different values and different application domains, and demonstrated to already take place in practice and specific cases. In short, the handbook presents Design for Values as a feasible possibility that enables important prospects to retain or regain moral control over technology, which can be demonstrated on the basis of existing and emerging practices.

The idea of making moral and societal values bear on technological developments can already be recognized in the effort of doing *technological assessments* (TA) (Grunwald 2009). These assessments were initially meant to identify upfront the outcomes of particular technological developments and aimed at providing governments in this way with a basis for judging the desirability of particular technology developments before they were adopted. This approach ran into the Collingridge dilemma, which states that early in the process of development of a technology, possibilities to intervene in the process are significant, but information about its unintended or undesired outcomes is scarce, while later on in the development of the technology, thanks to their technological momentum, steering of technology is either impossible, marginally possible, or prohibitively expensive (Collingridge 1980). Later forms of technological assessment such as Constructive Technology Assessment (Schot and Rip 1997) were therefore aimed at identifying and bringing in moral and societal considerations during the process of technology development, thus moving technology assessment toward including moral and social values in the design of technologies. They also addressed technology developers rather than the government, making technology assessment more anticipatory and proactive.

One of the historical origins of making social and moral values central to the design and development of new technology originated at Stanford in the 1970s, where it was a subject of study in computer science, and has been adopted by many research groups and is often referred to as Value Sensitive Design (VSD). Batya Friedman (Friedman 1997; Friedman and Kahn 2003; Friedman et al. 2006) was one of the first to formulate this idea of VSD; others have followed with similar approaches, e.g., "values in design" at University of California at Irvine (Bowker and Star 1999) and Values and Design at NYU (Nissenbaum 2001) and "design for values" at Delft University of Technology (van den Hoven 2007).

Finally, designers themselves have given moral and societal values a more central role over the course of time. Designers committed to technological

developments that are more meaningful from a societal point of view started to design for values and the concerns and needs of users. The approach of *participatory design* was launched as a method to include users in the design process and enabling to add their values to the design requirements and the artifacts that are to meet them (Schuler and Namioka 1993). Designers themselves thus think about incorporating sustainability, democracy, global development, and social improvement to their design requirements, as in *design for the base of the pyramid* and in *design for the real world* (Prahalad 2002; Papanek 1984). Moral and societal values are thus increasingly included in design, hence it is time to break away from the traditional views to see these values as externalities to design and explore the possibility, prospects, and practices of a deliberate Design for Values.

## The Prospects[1]

The prospects of Design for Values are obvious and tempting. It recognizes design as a far richer process since it can now be seen not only realizing our functional requirements but also our moral values. It recognizes designers as far more important professionals since they not only can provide us with technical means but can also address the values of people and society and think about expressing them in material culture and technology. Moreover, technological innovation is no longer a development that is separate from the values that users and society hold. Design for Values integrates design with our values and allows an active value-driven steering of and intervention in technological development. Design and designers can be frontloaded with moral and social values and are able to realize these values and can be held accountable for doing so. Technological innovation can become *responsible innovation*. The hope is that failure by societal opposition during implementation and adoption can be a phenomenon of the past as value issues are being addressed from the start. In this respect, Design for Values can contribute to the success, acceptance, and acceptability of innovations and as such will have also economic benefits.

The explicit and transparent articulation of values is thus important to innovation processes. These allow detecting dubious value commitments and allow designing for shared public values. The history of technology is full of examples where values have been obfuscated or tacitly lodged in designs or products. They range from "racist overpasses," which were designed to be so low as to prevent buses from poor black neighborhoods being routed to the beaches of the white middle class near New York (Winner 1980), to misleading biases in search engines, flaws in models in financial software – serious enough to "kill Wall street" – and deceptive maps in the user interfaces of geographic information systems (van den Hoven 2007).

Technical systems and innovative technology are the solidification of thousands of design decisions. Some of them were consciously taken after painstakingly

[1] Adopted from Van den Hoven (2013).

precise discussion among designers and engineers with good intentions. Some, however, were inserted negligently or malevolently to serve the interests of the designer or those commissioning him or her. What they have in common is that they may affect the lives of future users or entire societies. In the twenty-first century, we can help ourselves to the tools, methodologies, institutions, and procedures to discuss them explicitly and see to it that our world of technology and innovation is the best possible expression of our shared and public values.

This handbook gives an overview of the tools, methodologies, and procedures that have already been developed over the past few years. Despite the diversity in Design for Values approaches, we believe that they all have in common the following three characteristics.

First, there is the claim that values can be expressed and embedded in technology. Values can, for example, through their incorporation in technology, shape the space of action of future users, i.e., they can affect the set of affordances of and constraints to users. A road from A to B allows one to drive to B but not to C. Large concrete walls without doors make it necessary to take a detour. Architects and town planners have known this for quite some time. An ancient example not very different from the low-hanging overpasses of the early twentieth century is the so-called door of humility in the birth church of Jesus in Nazareth. The door is exceptionally low, and whoever wants to enter needs to bow his or her head, hence its name. The historical reason has been quite a different one from that of reminding people of the virtue of humility. The door was made intentionally low at the time of construction so as to make it impossible for mounted horsemen to enter the church on horseback in raiding attacks. If values can be imparted to technology and shape the space of actions of human beings, then we need to learn to incorporate and express shared values in the things we design and make.

Secondly, there is the claim that conscious and explicit thinking about the values that are imparted to our inventions is morally significant. Churchill famously observed, "first we shape our dwellings and then our dwellings start to shape us." Technology and innovation are formidable shapers of human lives and society. It is therefore very important to think about what we are doing to ourselves and to each other by means of technology. A good example is the recent work of Cass Sunstein entitled *Nudge*, which focuses on *choice architecture* (Thaler and Sunstein 2009). Think, for example, of the person who arranges the food in your university lunch room. By placing the deep-fried stuff almost beyond reach and the healthy fruit and veggies in front, the consumer is invited (not forced) to go for the healthy stuff (the nudge). Speed bumps and the "fly" in men's urinals are other examples of persuasion and nudging by technology. Sunstein and Thaler – following work in man–machine interaction research and cognitive ergonomics – provide many illustrations of how designers may arrange the feasible options of consumers and citizens in frivolous and serious contexts.

A third feature of any Design for Values approach is that value considerations need to be articulated early on in the process at the moment of the design and development when they can still make a difference. This sounds easier that it in fact is. This desideratum runs into the abovementioned Collingridge dilemma that

early in the process of technology development, possibilities to intervene are significant, but information about its outcomes is scarce, while steering later on in the development is either impossible or only marginally possible. The world of technology is a world of probabilities, ignorance, and uncertainty. Still Design for Values holds a claim that it is not only possible but also mandatory to assess value implications of design proactively. This is not to deny ignorance and uncertainty but rather requires designers to be more explicit about what they know and do not know and perhaps cannot know and to be more explicit about the value commitments they build into their designs so that they can assume accountability for their choices.

## The Practices

The practice of Design for Value is characterized by a diversity of approaches, theoretical backgrounds, values for which is designed, and application domains. For some values and in some application domains, Design for Values is a practice that takes place already for several years or even decades, and ample experience has now been accumulated and approaches established that have proven themselves. For other values and in other domains, the work just has begun. While there are approaches like Value Sensitive Design (VSD) that are influential for several values and application domains, practices and developments for different values and application domains are sometimes somewhat disconnected. This handbook brings these diverse experiences and approaches together and so provides a platform for further exchange and development.

The contributions gathered in Part I take a step back and describe the academic and historical sources of Design for Values. The main academic and historical source is VSD as already indicated, which originated in the domain of information technology. But also Technology Assessment, although initially not so much focused on design, is an important source, for example, in the health domain or in engineering. Participatory design also has developed over the years as a practice and approach that is an important source for Design for Values approaches.

Part II gives a more theoretical perspective with contributions about the possibilities and impossibilities of Design for Values and gives illustrations of extant practices. It discusses some general theoretical and methodological challenges and themes for Design for Values, like how to operationalize values in Design for Values so that they can guide the design process or how to deal with conflicting values. It also pays attention to design methods for Design for Values or how to model if one wants to design for values. Other theoretical contributions concern how to take into account the role of operators of sociotechnical systems in design, human capabilities, the role of emotions, and mediation, i.e., the way in which design affects the perceptions and actions of users.

The final two parts focus on practices of Design for Values in detail, in Part III for individual moral values ranging from accountability to trust and in Part IV for application domains ranging from architecture to water management. The parts

show how wide the range of values and application domains is in Design for Values but also lay bare the diversity of approaches and experiences. For some values like safety and more recently sustainability, Design for Values is now well established, and approaches are available that have proven themselves. For other values like responsibility, democracy, or justice, work is just beginning. Consequently, the contributions for the latter values are more exploratory in nature and describe possible approaches and applications rather than well-established approaches and experiences.

Something similar applies to the various application domains. Information and communication technology was the area in which the VSD approach developed, and Design for Values is in that domain now relatively well established. In other domains like water management, military technology, or nuclear energy technology, values may already play a role for quite some time, but they often do so implicitly rather than explicitly; no approaches for explicitly taking values into account are yet well established. In domains like economics and institutions, the challenges are again different.

To some extent, the diversity of tools and methods that the chapters in Parts III and IV show is a healthy sign reflecting the diverse challenges that different values and different application domains pose to Design for Values. A tool like life-cycle analysis that has proven its usefulness in design for sustainability may not be applicable to a value like responsibility. The military domain raises challenges for Design for Values like the fact that certain users or stakeholders may counteract the attempt to realize certain values that do not arise in other domains. At the same time, it seems clear that there are some common concerns and challenges like the need for design methods for Design for Values or ways to deal with conflicting values. Here, it would seem that the field can profit from more exchange between the different application domains and the different values. Hopefully, this handbook is a first step toward that exchange.

## References

Bowker GC, Star SL (1999) Sorting things out: classification and its consequences. MIT Press, Cambridge, MA

Collingridge D (1980) The social control of technology. Frances Pinter, London

Friedman B (ed) (1997) Human values and the design of computer technology. Cambridge University Press, Cambridge

Friedman B, Kahn PHJ (2003) Human values, ethics and design. In: Jacko J, Sears A (eds) Handbook of human-computer interaction. Lawrence Erlbaum, Mahwah, pp 1177–1201

Friedman B, Kahn PHJ, Borning A (2006) Value sensitive design and information systems. In: Zhang P, Galletta D (eds) Human-computer interaction in management information systems: foundations. Advances in management information systems, vol 5. M.E. Sharpe, Armonk, pp 348–372

Grunwald A (2009) Technology assessment: concepts and methods. In: Meijers A (ed) Handbook of the philosophy of science. Philosophy of technology and engineering sciences, vol 9. Elsevier, Oxford, pp 1103–1146

Nissenbaum H (2001) How computer systems embody values. IEEE Comput 34(3):118–120

Papanek VJ (1984) Design for the real world: human ecology and social change, 2nd edn. Van Nostrand Reinhold, New York
Prahalad CK (2002) The fortune at the bottom of the pyramid. Prentice Hall, New York
Schot J, Rip A (1997) The past and future of constructive technology assessment. Tech Forecast Soc Change 54(2/3):251–268
Schuler D, Namioka A (eds) (1993) Participatory design: principles and practices. Lawrence Erlbaum, Hillsdale
Thaler RH, Sunstein CR (2009) Nudge: improving decisions about health, wealth, and happiness, rev. and expanded edn. Penguin Books, New York
van den Hoven J (2007) ICT and value sensitive design. In: Goujon P, Lavelle S, Duquenoy P, Kimppa K, Laurent V (eds) The information society: innovation, legitimacy, ethics and democracy in honor of Professor Jacques Berleur S.J., vol 233. IFIP International Federation for Information Processing. Springer, Boston, pp 67–72. doi:10.1007/978-0-387-72381-5_8
Van den Hoven J (2013) Value sensitive design and responsible innovation. In: Owen R, Bessant J, Heintz M (eds) Responsible innovation. Wiley, Chichester, pp 75–84
Winner L (1980) Do artifacts have politics? Daedalus 109:121–136

# Part I

# Sources

# Value Sensitive Design: Applications, Adaptations, and Critiques

Janet Davis and Lisa P. Nathan

## Contents



**Abstract**

Value sensitive design (VSD) represents a pioneering endeavor to proactively consider human values throughout the process of technology design. The work is grounded by the belief that the products that we engage with strongly influence our lived experience and, in turn, our abilities to meet our aspirations. We, the

J. Davis (✉)
Computer Science, Grinnell College, Grinnell, IA, USA
e-mail: davisjan@cs.grinnell.edu

L.P. Nathan
University of British Columbia, Vancouver, BC, Canada
e-mail: lisa.nathan@ubc.ca



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_3

authors of this piece, are members of the first cohort of scholars to receive doctoral training from the founders of VSD at the University of Washington. We do not claim to represent an officially authorized account of VSD from the University of Washington's VSD lab. Rather, we present our informed opinions of what is compelling, provocative, and problematic about recent manifestations of VSD. We draw from contemporary case studies to argue for a condensed version of the VSD constellation of features. We also propose a set of heuristics crafted from the writings of the VSD lab, appropriations and critiques of VSD, and related scholarly work. We present these heuristics for those who wish to draw upon, refine, and improve values-oriented approaches in their endeavors and may or may not choose to follow the tenets of value sensitive design.

**Keywords**
Values • Human-computer interaction • Ethics • Stakeholders • Methodology

## Introduction

Value sensitive design represents a pioneering endeavor to proactively consider human values throughout the process of technology design. The work is grounded by the belief that the products that we engage with strongly influence our lived experience and, in turn, our abilities to meet our aspirations. Initially, the phrase "value sensitive design" was an umbrella term used to highlight an orientation towards human values shared between otherwise unaffiliated projects (Friedman 1999). However, since that time, value sensitive design (VSD) has become a branded term, designating specific strategies and techniques to help researchers and designers explicitly incorporate the consideration of human values into their work. To date, VSD has primarily been applied in the area of human-computer interaction (HCI).

Other branded, values-oriented approaches have developed in HCI, including Values at Play (Flanagan et al. 2005; Flanagan and Nissenbaum 2007), Values in Design (Detweiler et al. 2011; Knobel and Bowker 2011), and Worth-Centered Computing (Cockton 2009a, b). Participatory design has historically attended to participants' values (Iverson et al. 2010), although the term *values* is not always present. Still others are working in this area but do not present a branded account of how to do so (e.g., Flanagan et al. 2008), often focusing on specific values such as privacy (e.g., Palen and Dourish 2003; Barkhuus 2012). Within this field of endeavor, VSD is often recognized as the most extensive approach to date for addressing human values in technology design (Albrechtslund 2007; Le Dantec et al. 2009; Brey 2010; Fallman 2011; Manders-Huits 2011; Rode 2011; Yetim 2011).

We, the authors of this chapter, are members of the first cohort of scholars to receive doctoral training from the founders of VSD at the University of Washington. We were literally at the table as the approach was evolving.

We participated in the tangled, formative debates that rarely receive mention in formal academic writing. Because of this background, we offer a distinct perspective on recent methodological and theoretical applications of the approach and related critiques. We are able to identify authors with strong affiliations to the VSD lab, relationships that can be hard to discern from authorship and citations. No longer members of the VSD lab, we do not claim to represent an officially authorized account of VSD from the University of Washington's VSD lab. Rather, we present our informed opinions of what is compelling, provocative, and problematic about recent manifestations of VSD. Our envisioned readers are scholars who are (1) exploring this history and uptake of VSD as developed by Friedman and colleagues at the University of Washington, (2) interested in applying VSD in their own work, (3) working to extend or modify VSD, or (4) working in cognate areas.

We concentrate the majority of our analysis on the development of VSD since Friedman, Kahn, and Borning's seminal overview published in 2006. Friedman et al. (2006a) offer a thorough introduction to VSD and provide the first full articulation of what they term VSD's "constellation of features." The authors position this constellation as exclusive to value sensitive design (Friedman et al. 2006a). Taking this article and its claims as a point of departure, we examine how VSD has been appropriated and critiqued since the 2006 article was published. We draw from contemporary case studies to argue for a condensed version of the VSD constellation of features. We also propose a set of heuristics crafted from the writings of the VSD lab, appropriations and critiques of VSD, and related scholarly work. We present these heuristics for those who wish to draw upon, refine, and improve values-oriented approaches in their endeavors and may or may not choose to follow the tenets of value sensitive design.

## Method for Collecting Articles for Review

The scholarship discussed in this chapter is primarily VSD-influenced research and design from the years 2007–2012. We began our search for related work in the ACM Digital Library. We proceeded to expand the search to databases and journals from cognate fields (information and library science). Specific search terms and data parameters are provided in Table 1. We removed from our analysis writings that were (1) conference workshop proposals or panel position papers, (2) magazine articles discussing designers' practice around values that do not explicitly address the values-oriented design scholarship, (3) pedagogical work, or (4) unpublished works in progress. We added some works not identified in our search, but cited by sources identified in our search.

This chapter is not an exhaustive review of all work that has incorporated, cultivated, or critiqued VSD in the past 6 years. Our goal is to present scholarship that exemplifies the development of VSD to date. There is worthwhile work that we did not examine in this chapter.

**Table 1** A summary of databases and search terms used in our search for recent VSD-influenced research

| Databases/Journals | Search terms | Time periods |
|---|---|---|
| ACM digital library | "Value sensitive design"; value sensitive design; VSD | 2007–2012 |
| Library and Information Science Abstracts (LISA) | "Value sensitive design"; value sensitive design | 2007–2012 |
| Library, Information Science & Technology Abstracts (LISTA) | Value sensitive design; value and sensitive and design | All |
| Google scholar | Value sensitive design | 2007–2012 |
| Journal of ethics and information technology | Value sensitive design; VSD | All |
| Journal of information science and technology | "Value sensitive design"; value sensitive design; VSD | All |

## History of Value Sensitive Design

Under explicit development since early in the 1990s, VSD is claimed to provide a theory (Friedman and Freier 2005), a methodology (Friedman and Kahn 2003; Friedman 2004), methods (Miller et al. 2007; Nathan et al. 2007), and an approach (Nathan et al. 2008; Woelfer and Hendry 2009) for scaffolding consideration of human values during the design, implementation, use, and evaluation of interactive systems. An early explication of VSD is found within a National Science Foundation (NSF) workshop report (Friedman 1999). The report uses "value sensitive design" as a label for a wide range of projects undertaken by scholars who were likely unaware of the term, but whose interactive design work shared a proactive approach to addressing concerns other than efficiency and usability (Friedman 1999). Soon thereafter, Batya Friedman and a core group of collaborators at the University of Washington began publishing research describing and practicing what they formally termed "value sensitive design" in books, journals, conference proceedings, and white papers (http://www.vsdesign.org/publications.php). Over time, work cited as representative of VSD (Borning and Muller 2012) typically is coauthored by Friedman or other researchers affiliated with the VSD Lab, suggesting a proprietary relationship between the VSD approach and the VSD Lab.

What is a value according to VSD? Friedman's early explications of VSD did not define the term "value" explicitly, but instead listed broad areas of concern including human dignity and welfare (Friedman 1999). She proceeds to highlight certain values deserving of attention: "trust, accountability, freedom from bias, access, autonomy, privacy, and consent" (Friedman 1999, p. 3). In the 2006 article, the term value was defined as "what a person or group of people consider important in life" (Friedman et al. 2006a, p. 349). However, this rather broad definition was circumscribed in part by another list of specific values "with ethical import that are often implicated in system design" (Friedman et al. 2006a, p. 349).

Through our review, we found no statements that VSD is a finished product; rather, we found it offered for others to continue to adapt and improve (Borning and Muller 2012). To position a discussion of how VSD has been used and critiqued, the following paragraphs briefly describe each of the areas mentioned above: theory, methodology, method, and approach.[1]

## Theory

Key to value sensitive design is its basis in an interactional understanding of technological appropriation. This theoretical positioning claims that a technology's influence on humanity is shaped by the features of its design, the context in which it is used, and the people involved in its use. To ignore any component of this emergent and relational process (tool features, context, or stakeholders) is problematic. The interactional perspective implies that the impact of a technology on human life is not fully determined by the technology's design. Values are not embedded within a technology; rather, they are implicated through engagement. A technology can be appropriated in innumerable ways, shaped by individuals and societies and by the context of use, as well as by its form and content.

VSD collaborators and allies believe that a concerted effort to identify and address human values implicated by technology use – even though that effort is imperfect – can significantly improve the design of products. In turn, when we ignore the influence of a product's use on lived experience, the resulting interactions are more likely to have a range of unintended, negative impacts on human lives (Nathan et al. 2007). Moreover, as we discuss later, the effects of interactions with technology reach far beyond those who are directly involved in technology use.

## Methodology

Early VSD literature emphasized the development of a methodology for addressing values. This "tripartite" methodology is composed of three types of iterative and integrative "investigations," labeled conceptual, empirical, and technical (Friedman and Kahn 2003).

*Conceptual investigations* involve two primary activities. The first is identifying the stakeholders who will be affected by the technology under study. This includes those who use (or will use) a given product (*direct stakeholders*) and those who may not engage the technology directly, but whose lives will be influenced through others' use (*indirect stakeholders*). As an example, a conceptual investigation of a building surveillance system would likely identify the security personnel who manipulate, maintain, and monitor the system as direct stakeholders. The indirect

[1]For a more detailed account of VSD, we recommend Friedman et al. (2006a) overview.

stakeholders might include building inhabitants and visitors (welcome and unwelcome) whose images will be captured by the camera. Although these latter individuals do not directly interact with the system, their lives are influenced by others' use of the technology.

The second component of a conceptual investigation is identifying and defining the *values* implicated by use of a technology. For example, conceptual investigations of building security technology in a condominium in Vancouver, Canada, would include creating definitions of what values such as privacy mean in that context, where privacy is a legislated right. Value conflicts (or tensions) can emerge as soon as values are identified and discussed (Friedman et al. 2006a). In this example, residents' conceptualizations labeled *security* regarding their personal safety and possessions might stand in tension with expectations related to the *privacy* of residents and nonresidents who enter the building.

*Empirical investigations* examine stakeholders' "understandings, contexts, and experiences" in relation to technologies and implicated values (Friedman and Kahn 2003). Such investigations may employ a variety of methods – surveys, questionnaires, interviews, experiments, artifact analysis, participant observation, and so on – to inquire into stakeholders' observable actions as well as their understandings, concerns, reflections, and aspirations.

*Technical investigations* are primarily concerned with specific features of technologies. These studies may include designing a new technology to support particular values or analyzing how particular features of existing technologies implicate certain values in a context of use.

It is worth reiterating that the investigations are presented as iterative and integrative (Friedman et al. 2006a). They are meant to inform each other rather than be engaged as separate, modular activities. Investigations may overlap, happen in different orders, or intertwine with each other. One activity can serve multiple purposes. Even researchers working with the VSD lab who call out VSD as a primary influence do not always elucidate these three investigations (e.g., Friedman and Hendry 2012).

## Methods

Since its inception, VSD has incorporated the investigation of values into a range of standard social science methods, such as semi-structured interviews (Friedman et al. 2006a), surveys, observations, quasi-experimental designs, exploratory inquiries (Woelfer et al. 2008), and longitudinal case studies (Nathan et al. 2009). Researchers associated with the VSD lab in Seattle have taken a values orientation in their use of physiological measures (e.g., Kahn et al. 2008) and chat log analyses (Friedman et al. 2003; Kahn et al. 2005), as well as design methods such as probes (Nathan et al. 2009), sketching (Woelfer et al. 2011), and scenarios (Nathan et al. 2008; Woelfer and Hendry 2012).

Friedman recently laid claim to the development of 14 unique value sensitive design methods (Friedman 2012). The list consisted of (1) stakeholder analysis;

(2) designer/stakeholder explicitly supported values; (3) coevolution of technology and social structure; (4) value scenarios; (5) value sketches; (6) value-oriented semi-structured interview; (7) granular assessments of magnitude, scale, and proximity; (8) value-oriented coding manual; (9) value-oriented mock-ups, prototypes, and field deployments; (10) ethnography focused on values and technology; (11) model for informed consent online; (12) value dams and flows; (13) value sensitive action-reflection model; and (14) envisioning cards. We were unable to find explicit descriptions of all of these in the published literature on VSD. Some (3, 4, 5, 6, 9, 10) are examples of values-oriented appropriations of social science and design methods mentioned in the preceding paragraph. Below, we discuss three methods from this list that are reported upon in the literature: direct and indirect stakeholder analysis, value dams and flows, and the envisioning criteria and cards. The first two can be positioned as values-oriented analyses; the last is a values-oriented toolkit.

*Direct and Indirect Stakeholder Analyses*. In a stakeholder analysis, researchers attempt to identify the roles of individuals who will be affected by the technology under study. This includes those who use (or will use) a given technology (*direct stakeholders*) and those who may not engage the technology directly, but whose lives will be influenced through others' use of the technology (*indirect stakeholders*). These roles might be distinguished by job type (e.g., programmer, conductor), relation type (mother, daughter), interaction with technologies (e.g., contributor, reader, commenter), or any of the other myriad positions that individuals take on in daily life. The types of roles will depend on the context(s) under investigation.

Note that the term stakeholders refers to *roles* and not individual people. For example, we (Janet and Lisa) are not solely defined by our roles as authors of this chapter. Depending on the context, we may also be described as teachers, daughters, friends, women, citizens, voters, bicyclists, or gardeners. As we engage with a tool throughout a day, we take on any number of these different stakeholder roles, depending on the situation. Conceptualizing an individual as *mother* rather than as in the role of *mother* risks ignoring the multiplicity of roles through which we engage with our environments (Friedman et al. 2006b; Miller et al. 2007).

In a robust stakeholder analysis, researchers consider both the roles of those who will use a tool along with the roles of those who will be affected by others' use of the tool. For each role, the research team identifies the potential harms and benefits that these roles are likely to experience as a result of the tool under investigation being used.

Note that stakeholder analysis is not unique to VSD. Such analysis is also found in fields such as public policy, conflict resolution, and business administration. However, in these other areas, the concept of indirect stakeholder may not be present. In VSD, the goal of the stakeholder analysis is to iteratively inform further conceptual investigations focused on values, as well as empirical and/or technical investigations, by framing the roles that should be considered (Friedman et al. 2006a; Miller et al. 2007).

*Value Dams and Flows*. The value dams and flows method is a process for making decisions regarding value tensions (Miller et al. 2007). Value tensions

occur when conceptualizations of values or the design implications of values are found to be in friction with each other. Value tensions can surface in a variety of ways. For one, values articulated by an individual may conflict with those shared by a group. A familiar example of this type of tension can be found in information practices that develop around the use of organizational open calendaring systems. An open calendaring system is purported to support collaboration and accountability within a group. However, having one's daily schedule available for all to see can also be perceived as intrusive. As a result, information practices may develop to address the varied tensions between collaboration, accountability, privacy, autonomy, and identity (Palen 1999). A second type of tension can occur when a system supports a societal value that runs counter to an organizational value. An example of this type of tension is an interactive system that supports collaboration (in this case a societal value), deployed in an organization that explicitly rewards fierce individualism and competition (Orlikowski 1992). Value tensions can also occur within individuals or across groups.

VSD does not claim that all value tensions can (or even should) be resolved. Once a tension has been identified, one may choose to address the tension or perhaps mark it for attention at another stage in the process (Miller et al. 2007; Nathan et al. 2007).

In the value dams and flows method, value dams are defined as technical features or organizational policies that some stakeholders (even just a few) strongly oppose, causing a value tension. The proposition is that it is important to identify potential opposition to how a tool functions (features) or is deployed (policies) because strong opposition by even a few can block appropriation of the technology. If the findings from a survey or set of interviews suggest that some stakeholders hold a strongly negative view of a technological feature or policy, it is likely that a value dam will develop, inhibiting effective use of the technology. Value flows are the flip side of this construct, calling attention to features and policies that a large number of stakeholders are in favor of incorporating. Attending to value flows in the design of tool features and policies may attract people to adopt the tool or to intensify their use (Miller et al. 2007).

*Envisioning Criteria and Cards*. The consideration of values when working with an interactive technology is complex and hard (Fallman 2011). Undertaking such work may appear out of scope when designers are already pushed to meet numerous commitments with limited time and too few resources. To address this challenge, members of the VSD lab looked to the scholarship in urban planning and design noir for insights on how to engage longer-term thinking in complex environments as well as address the reality that designs are often appropriated in unforeseen ways. From this multidisciplinary inquiry developed four "envisioning criteria": stakeholders, time, values, and pervasiveness (Nathan et al. 2008). In turn, these four criteria were expanded into the Envisioning Cards toolkit, a product meant to support agile consideration of values during the design, redesign, or implementation of an interactive system (Friedman and Hendry 2012).

Each set of Envisioning Cards includes a stack of 32, $3'' \times 5''$ cards, a sand timer, and a small information booklet. Four of the cards are blank, encouraging users to

create new, context-specific cards. Each of the other 28 highlights one of the four envisioning criteria (stakeholders, time, values, or pervasiveness), a card theme, and an activity on one side. The reverse side has an image that is meant to evoke or support that card's particular theme. Just as the order of VSD investigations is intentionally flexible, the Envisioning Cards are designed to provide opportunities for values-oriented reflection, iteration, and course correction throughout the design, implementation, and evaluation of information tools (Friedman and Hendry 2012). To date, documented uses of the cards are limited to supporting classroom activities, projects with members who are associated with the VSD lab in Seattle, and conference workshops (Kaptein et al. 2011; Friedman and Hendry 2012).

## Approach: Variations in VSD Uptake

Since 2006, VSD's influence has become increasingly apparent in publications beyond those of Friedman's group at the University of Washington.[2] Some of these appropriations have used the tripartite methodology, others the stakeholder analysis, and still others simply an orientation towards values where VSD's influence is mentioned in the literature review. We are not interested in demarking who is or who is not doing *straight-up* value sensitive design research. However, we believe that these variations in practice, different approaches to VSD, are important to identify and may lead to future lines of inquiry.

Here we attempt to describe the ways in which recent scholarship has related to VSD. We consider four categories: an affinity with VSD, prescriptive VSD, critical VSD, and formative VSD. These categories do not neatly divide the space; rather, these categories have fuzzy boundaries and regions of overlap. The works we point to within these categories are offered as examples.

*An affinity with VSD*. The values-oriented research area in the field of human-computer interaction has continued to evolve since Friedman's 1999 NSF report (e.g., Blevis 2007; Flanagan et al. 2005; Klasnja et al. 2009; Le Dantec et al. 2009; Palen and Dourish 2003). Friedman's lab at the University of Washington is just one strand.[3] The affinity among these approaches grows from a shared concern regarding the complex interplay between human values and the design and use of information tools. The information tools under consideration are not always new and networked; there also investigations of non-digital tools (e.g., Woelfer and Hendry 2009; Wyche et al. 2007; Wyeth 2006).

Such work may explicitly critique VSD (e.g., Le Dantec et al. 2009; Leitner et al. 2010) or draw inspiration from VSD (e.g., Chango 2007; Foong 2008), without claiming an adherence to VSD theory or methodology. Even when

[2]We did not find any published work describing VSD projects prior to 2006 that did not involve researchers with strong links to Batya Friedman and the development of VSD (e.g., Helen Nissenbaum, Peter Kahn, Alan Borning).

[3]However, it is a strand that has become a brand.

values-oriented work does not claim any particular relationship to VSD, it often shares key features such as an interactional perspective and a proactive approach to values in design. Notably, Flanagan et al. (2008) describe technical, philosophical, and empirical modes of inquiry, aligning with VSD's technical, conceptual, and empirical investigations. This similarity may reflect ideas developed in prior collaborations between Friedman and Nissenbaum (1996, 1997).

*Prescriptive VSD*. Many researchers and designers have taken VSD at face value, applying it as presented by Friedman and colleagues, often in a new domain. Some work applies VSD to technology design in a particular context of use, for example, supporting healthcare workers in Africa (Walton and DeRenzi 2009) and blind transit users in the United States (Azencot et al. 2011). Other work aims to establish principles or frameworks that account for ethics in designing particular technologies, such as brain-computer interfaces (Simbasavian and Jackson 2007), persuasive technology (Davis 2009), social networks (Cotler and Rizzo 2010), healthcare robots (van Wynsberghe 2013), and nanopharmaceuticals (Timmermans et al. 2011).

*Critical VSD*. Other researchers have critiqued VSD from a variety of perspectives. Some have applied VSD and found that it falls short. For example, Johri and Nair (2011) apply VSD to an e-governance project in India. In their discussion, they observe that "there were several issues that were not as clear-cut in the field as noted in the framework" – in particular, contextual and emergent aspects of values, the importance of pragmatic issues, unresolved contradictions between values, and the role of intermediation in access to technology. Others critique VSD in explaining why they choose another approach (e.g., Le Dantec et al. 2009; Halloran et al. 2009). Still others critique VSD in order to develop or extend the. We consider such critiques of VSD at greater length in the next section.

*Formative VSD*. Much work inside and outside the VSD lab aims to develop or extend VSD.

While Borning and Muller (2012) aim to improve the overall practice of VSD and increase its uptake, others aim to fill gaps in VSD theory or in the methods offered. For example, Yetim (2011) argues that VSD suffers from the lack of an ethical theory and shows how discourse ethics can fill this gap, notably offering guidance for stakeholder analysis. Pommeranz et al. (2011) argue, after Le Dantec et al. (2009), that VSD empirical investigations need methods for situated elicitation of values; they compare the photo elicitation interview with other methods and propose a new value elicitation tool. Detweiler et al. (2011) argue that VSD lacks formal modeling methods to inform technical investigations of values; they demonstrate how to incorporate values into the Tropos approach to requirements modeling.

## Critiques

As awareness of VSD has grown, so have the numbers of critiques. Some of these critiques are explicit, calling out problematic areas of VSD in terms of how it has been formulated, presented, and applied (Le Dantec et al. 2009; Borning and Muller 2012). Other critiques refer generally to values-oriented design approaches, carefully

avoiding particular labels, but offering relevant insights (Bardzell 2010). As a body of scholarship, we find these critiques offer some of the most stimulating and provocative work in this area over the past 6 years. Here we engage the critiques that address core aspects of VSD: universal values, ethical commitments, stakeholder participation, the emergence of values, and the voices of researchers and participants.

## Universal Values

A frequent critique of VSD concerns its position that certain values are universal, although those values may play out in different ways across cultures and contexts (Friedman et al. 2006a). Borning and Muller (2012) argue that this position sits atop a slippery slope: "The belief that there are universal values . . . has on occasion led to the further belief that a particular group, culture, or religion is the keeper of those values, and needs to impose them on others – with sometimes tragic consequences." Borning and Muller (2012) further claim that VSD's stated commitment to universal values has likely impeded uptake of the approach. Indeed, others have explicitly or implicitly critiqued this commitment (e.g., Le Dantec et al. 2009; Alsheikh et al. 2011), and even some who claim VSD as their approach reject VSD's stance on universal values (Johri and Nair 2011).

Borning and Muller (2012) propose two different responses to such a quandary: to shift from a philosophical to an empirical basis for one's stance or to make explicit the researcher's position, implicitly acknowledging that theirs is not the only valid position. As Borning and Muller (2012) point out, the founders of VSD claim that their position is supported by empirical evidence (Friedman and Kahn 2003, Friedman et al. 2006a). Yet, this position is still vulnerable to critique, for systematic abuses of the unprivileged have often been justified on "scientific" grounds (Borning and Muller 2012). Moreover, empirical positions can be overturned by contradictory evidence. For example, Hofstede's (1991) empirically based model of cultural values identifies dimensions along which residents of different nations espouse opposing values: for example, individualism versus collectivism. Where Friedman et al. (2006a) claim privacy and autonomy as universal values (although their expression differs across different cultures), Saab (2008) instead casts these values as belonging to individualist cultures, in contrast to values such as group cohesion held in collectivist cultures. However, Saab (2008) also notes a need for further empirical validation of this ethnorelativist model, particularly in contrast to VSD's universalist stance. Thus, a universalist position based on empirical evidence remains a contentious position.

Turning to the second response, Friedman et al. (2006a) certainly make their position explicit, but do so by claiming that a commitment to universal values is obligatory for those practicing VSD. Does VSD require universal values as part of its foundation? While arguing strongly that particular values are universal – specifically, human welfare and ownership of property – Friedman and Kahn (2003) acknowledge with Saab (2008) that not all values are universal and some cultures hold contrasting values, such as cooperation versus competition.

Furthermore, Friedman and Kahn (2003) go on to address the implications of cultural variability for design: when implicated values vary across cultures, systems may be appropriate only within a particular culture, unless designers make an extra effort to build in value adaptivity. Some variability is thus already accounted for in VSD, but there is not agreement on whether it provides enough flexibility.

Building on these ideas, Borning and Muller (2012) argue for a pluralistic stance: that VSD should not recommend any position on the universality or relativism of values, but rather leave VSD researchers and practitioners free to take and support their own positions in the context of particular projects.

## Ethical Commitments

While some critique VSD for its adamant commitment to the general concept of universal values, others critique VSD for failing to make concrete ethical commitments. While praising VSD for addressing universal values and drawing on ethical theory, Albrechtslund (2007) points out that VSD leaves unclear "*what* values and *which* theories" it includes. Without an explicit commitment to an ethical theory, he claims, VSD is an ethically neutral tool, vulnerable to use in support of harmful values such as those of Nazism.

Indeed, descriptions of VSD do not recommend the use of any particular ethical theory. Rather, Friedman and Kahn (2003) argue that values relevant to information technology design find their basis in different kinds of ethical theories. Some values – such as not intentionally deceiving others – rest on theories of the right (consequentialist and deontological ethics), which concern moral obligations and prohibitions. Other values such as "warmth and friendliness," they argue, would not be considered moral values under such theories, as there is no obligation to be friendly. But from the perspective of virtue ethics, or theories of the good, these are indeed moral values, as one is a better person for being warm and friendly (Friedman and Kahn 2003). Friedman and Kahn (2003) imply that designers *must* attend to values supported by theories of the right, which are obligatory, and *may* attend to values supported by theories of the good, which are discretionary. However, they make no commitments to particular theories.

Manders-Huits (2011) advances this critique, arguing that the notion of values in VSD is "underdeveloped." She claims that VSD provides "no methodological account for distinguishing genuine moral values from mere preferences, wishes, and whims of those involved in the design process" (Manders-Huits 2011). That is, among the many things that stakeholders consider important in life, how does the investigator determine which ones correspond to values of ethical import that *ought* to be attended to in design? Without such an account, the VSD practitioner risks attending to a set of values that is unprincipled or unbounded.[4] The solution,

[4]Indeed, we, the authors, have questioned the grounds on which Friedman et al. (2006a) exclude usability from the realm of moral values while including values such as calmness and courtesy (Friedman et al. 2006c, p. 4).

Manders-Huits (2011) argues, is that VSD requires a complementary ethical theory not only to demarcate moral values but also to provide a basis on which to make principled judgments about which values are most important to support. In extending VSD to "value conscious design," Manders-Huits (2011) recommends that designers clarify their ethical goals and explicate their chosen ethical theory.

Several VSD projects have adopted ethical theories found suitable to the project domain. For example, Borning et al. (2005) draw upon discourse ethics to support the legitimation of an urban simulation, while Chatterjee et al. (2009) explicitly adopt deontological ethics in the context of developing collaboration systems. Writing about the design of a hypothetical weapons command and control system, Cummings (2006) draws on the theory of just war and the principles of proportionality and discrimination. Similarly, in developing a VSD-based framework for the ethical design of healthcare robots, van Wynsberghe (2013) adopts the perspective of care ethics, which emphasizes relationships and responsibilities over rights. Thus, the literature provides several models for following Manders-Huits' (2011) recommendations to explicitly adopt an ethical theory alongside the VSD approach.

By contrast, Yetim (2011) argues that discourse ethics is a uniquely appropriate standpoint from which to critically examine VSD itself, going beyond any particular project domain such as urban simulation. Even though Borning et al. (2005) address the legitimacy of the simulation software and the transparency of the software development process, Yetim (2011) argues that they do not go far enough in addressing the legitimacy of the design process itself. Yetim (2011) therefore develops a *general* approach for adopting discourse ethics alongside VSD, in support of the legitimacy of the value sensitive design process, while at the same time acknowledging that discursive methods may play a different role in different design contexts.

## Stakeholder Participation and the Emergence of Values

Turning the lens of discourse ethics to VSD itself provides Yetim (2011) with theoretical grounding for a critique of the role of stakeholders in VSD work. According to Yetim (2011), "the discourse principle suggests the inclusion of all those affected in discourse, which in turn requires a method to identify them." This raises two issues. First, along with Manders-Huits (2011), Yetim (2011) claims that VSD fails to provide a systematic and comprehensive method for identifying stakeholders. Second, Yetim (2011) argues that VSD fails to address the use of deliberative methods and tools to promote joint reflection on values during the design process – in particular, reflection by stakeholders on their own values, value tensions, and implications for design, as participants in the design process.

Others, too, have called for greater stakeholder participation in the VSD process. For example, Kujala and Väänänen-Vainio-Mattila (2008) emphasize the need for stakeholders to reflect upon their own values, as do Pommeranz et al. (2011). Borning and Muller (2012) further argue that stakeholders should have a greater voice in the VSD process. They observe that in recent years participatory design (PD) has

extended far beyond the workplace and workers; indeed, some recent PD work has aimed to "rekindle" attention to values in participatory design (e.g., Iverson et al. 2010; Halloran et al. 2009). Borning and Muller (2012) go on to recommend that VSD projects consider explicit commitments to codesign and power sharing.

Bolstering this call for greater stakeholder participation is a critique of VSD's systematic approach to identifying values of concern. Le Dantec et al. (2009) are concerned that Friedman and Kahn's (2003) list of "12 human values with ethical import" serves to reify the values already studied in the HCI community and further privilege them over the values held by stakeholders, which might be quite different. Although one of the purposes of empirical investigations is to serve as a check on conceptual investigations, Le Dantec et al. (2009) argue that having a list of values may blind the researcher to values that fall outside that list. Rather, they promote what Iverson et al. (2010) call an *emergent* approach to values, where the values at stake initially emerge from work with stakeholders rather than an initial conceptual investigation carried out by the researchers alone.

While still calling for greater stakeholder participation as noted earlier, Borning and Muller (2012) soften this critique, observing that VSD has evolved from being highly prescriptive in listing values worthy of concern (Friedman and Kahn 2003) to providing suggestive heuristics as in the Envisioning Cards (Friedman and Hendry 2012). Moreover, Borning and Muller (2012) argue that such heuristics can be useful: heuristics enable projects where VSD would otherwise be impractical to build on values-oriented work in the literature. Even when there is adequate time for empirical investigations, heuristics may reduce, rather than increase, the risk that designers will overlook areas of concern. At the same time, Borning and Muller (2012) caution that heuristics should be contextualized, recognizing who developed the heuristics (in this case, Western, upper-middle class academics).

Out of these critiques come methods for advancing the application of VSD. Yetim (2011) recommends the use of Ulrich's (2000) "critically heuristic boundary questions" to identify stakeholders. These questions concern sources of motivation, power, knowledge, and legitimation. Yetim (2011) suggests that the questions be addressed iteratively, in both descriptive and prescriptive modes, to uncover unresolved issues.

Recent work demonstrates a spectrum of methods concerning emergent values and participation, varying from value elicitation activities in which researchers make meaning from observations of stakeholders to participatory approaches where researchers and stakeholders together create shared meanings:

- Alsheikh et al. (2011) use ethnography to defamiliarize their own values in a cross-cultural setting; they use grounded theory to identify themes in their observations.
- Woelfer et al. (2011) elicited participants' understanding of the value *safety* through value sketches and scenarios created by the participants and interpreted by the researchers.
- Le Dantec et al. (2009) recommend the photo elicitation interview (PEI) for "shifting the power dynamic towards the participants by letting them shape the

direction of the interview." Pommeranz et al. (2011) compare the PEI with the portrait value questionnaire (Schwartz and Bilsky 1990) and participant tagging of photographs, finding that the PEI gives more descriptive and more situated values, but still fails to elicit how values inform behavior, decisions, and trade-offs. They suggest the development of a mobile app for in situ elicitation of values.

- Iverson et al. (2010) discuss a variety of methods for discovering, developing, grounding, and realizing values in collaboration with stakeholders. Halloran et al. (2009) show how participatory design workshops and interactive prototypes can elicit values, observing that "values [emerge] whether or not we look for them."

In contrast, Yetim (2011) problematizes value discovery as part of design: it is infeasible to include all stakeholders in discourse about values, and interpretations of values and tools may change over time. He points out that UrbanSim is designed for technical flexibility so that developers can respond to concerns that emerge during the use of the system (Borning et al. 2005), yet UrbanSim includes no tools or process for eliciting those emergent concerns. In response, Yetim (2011) recommends "continuous participation and discourse": systems should include tools for communication about breakdowns in the system itself. In particular, his DISCOURSIUM tool (Yetim 2008) draws on discourse ethics to structure reflection on comprehensibility, relevance, validity, and rationality.

## Voice

Borning and Muller (2012) present two compelling critiques related to VSD and issues of voice. First, Borning and Muller (2012) call for greater attention to the voice of the researcher. They claim that, too often, VSD research is reported from a disembodied "we" position, with the authors failing to clarify who is making various claims throughout the work. Borning and Muller are not asking for the researchers to simply claim ownership of their statements, but to help the reader understand the researchers' backgrounds and values. Do the researchers and the stakeholders they are reporting on have similar backgrounds? Are there conflicts or tensions in how stakeholder and researchers view the situation under study? Borning and Muller call attention to the influence of researchers' perspective on what they find important in an investigation. A researcher is not a disembodied conduit for truth, but rather takes an active role in interpreting, analyzing, and designing. Strong examples of making the researchers' position explicit include the scholarship of Ames et al. exploring the role of social class in technological appropriation (Ames et al. 2011), and a growing body of work by Woelfer and Hendry on information technology for homeless youth (particularly Woelfer et al. 2011). Both groups of authors make clear statements about their positions as researchers.

Second, Borning and Muller (2012) raise concerns regarding the voice of stakeholders and how the multiplicity of voices is identified, brought forward,

and attended to throughout the lifecycle of a project. As mentioned in the previous section, Borning and Muller recommend that VSD scholars consider stakeholder participation and voice throughout the entire research process. Beyond issues of participation, Borning and Muller address the presentation of research, arguing that summaries and paraphrases place researchers at risk of unintentionally reporting their own values or thoughts as if they were the values or thoughts of the participants. The usual response to this problem in HCI and other fields is to liberally use direct quotations from participants in final publications. This provides readers with the opportunity (albeit imperfect) to engage directly with the stakeholder's choice of words, their own voice. Several examples of this practice can be found in research on values and technology (Alsheikh et al. 2011; Czeskis et al. 2011; Fleischmann et al. 2011; Woelfer and Hendry 2012).

## VSD Looking Forward: Commitments and Heuristics

As mentioned earlier, in 2006 Friedman, Kahn, and Borning identified a "constellation" of eight features distinguishing VSD from other approaches to design (Friedman et al. 2006a). Here, we reposition the constellation, paring it down to commitments that can guide those engaging in a VSD investigation and that are largely uncontested[5] in the literature. Moreover, from the range of VSD appropriations, we draw out general heuristics for individuals embarking on values-oriented projects, who may or may not wish to position themselves as engaged in value sensitive design.

### Core Commitments

Drawing upon the structure used by Friedman et al. (2006a), we identify four core commitments of VSD: proactive stance, interactional perspective, direct and indirect stakeholders, and tripartite methodology. We illustrate these commitments through recent case studies. We do not mean to imply that all VSD work rests equally on these four commitments, nor do we intend to draw a sharp line between research that adheres to each of these commitments and work that does not. Rather, we aim to demonstrate how these core commitments make unique contributions to a range of VSD manifestations.

[5]Regarding the other four features listed in the 2006 article, we claim that three help distinguish VSD from other approaches, but are neither unique to VSD nor essential for this scholarship: (1) VSD enlarges the design arena beyond the work place, (2) VSD enlarges the scope of human values beyond those of cooperation and participation and democracy, (3) VSD distinguishes between usability and human values with ethical import. (4) The last feature, adherence to the proposition that some values are universal, is a contested position (Borning and Muller 2012) addressed earlier in this chapter.

*Proactive stance*. VSD is proactive in two ways. First, it positions researchers to proactively identify ethical concerns implicated by interactions with and through technology rather than waiting for an ethical problem to arise. For example, van Wynsberghe (2013) considers the nascent field of healthcare robotics, articulating the need for a framework that incorporates ethics and design. In the absence of universal guidelines or standards for robot design, she recommends the adoption of VSD in combination with a care ethics perspective. van Wynsberghe (2013) points out that VSD can be used both retrospectively, to analyze the ethical implications of existing care robots (regardless of whether problems have already occurred), and proactively, to guide consideration of ethics throughout the process of designing care robots.

Second, VSD "seeks to be proactive: to influence the design of technology early in and throughout the design process" (Friedman et al. 2006a). While this proactive stance is not unique to VSD, it is an essential feature that distinguishes a *design* approach from critique or analysis of existing technologies. This stance draws attention to values both *early* in the design process and *throughout* the design process. As with privacy, security, or usability, support for values cannot always be "bolted on" late in the design process, but rather requires that designers make fundamental decisions about requirements and architecture with those values in mind. Furthermore, key values should be not be forgotten in the face of competing concerns, but rather reconsidered at each step of design and evaluation.

A proactive approach to values can make a difference in outcomes. Davis (2009) compares two contemporary projects with similar goals, one that takes a VSD approach and one that does not. Both projects aim to develop tools that facilitate knowledge sharing within an organization. BlueReach is designed from the perspective of persuasive technology (Brodie et al. 2007), while CodeCOOP is designed using VSD theory and methodology (Miller et al. 2007). Although both consider the value of *reputation*, the persuasive technology perspective considers reputation only as a strategy to promote knowledge sharing. The assumption is that an individual enhances her reputation through publicly sharing useful information (Brodie et al. 2007). In contrast, the VSD perspective led Miller et al. (2007) to also consider potential harms to reputation (e.g., from asking a silly question) that might impede use of the system. Moreover, Miller et al. (2007) considered a more expansive field of values from the start of the design process, including privacy, trust, and awareness, along with reputation. Early empirical investigations positioned the CodeCOOP designers to assess and mitigate these value tensions before building the CodeCOOP system, thereby leading to an apparently successful deployment (Miller et al. 2007). In contrast, empirical investigations of barriers to BlueReach's use took place after a less-than-successful deployment; only then did harms to reputation emerge as a concern that stopped people from using the system (Singley et al. 2008). Thus, VSD's proactive stance guided Miller et al. (2007) to address users' concerns early in the design process, avoiding potential barriers to system adoption.

*Interactional perspective*. VSD takes an interactional perspective: "values are viewed neither as inscribed into technology (an endogenous theory) nor as simply

transmitted by social forces (an exogenous theory). Rather, the interactional position holds that while the features or properties that people design into technologies more readily support certain values and hinder others, the technology's actual use depends on the goals of the people interacting with it" (Friedman et al. 2006a). For the designer to hold an exogenous theory is a defeatist position: What is the point in designing for values if technology has no influence on how values are expressed? But neither should the designer adhere to an endogenous theory, which would overclaim the designer's ability to determine which values the technology implicates and how values will ultimately be expressed. Falling between these two extremes, the interactional position is a widely accepted theoretical stance on socio-technical systems that can productively guide design.

Building on this interactional perspective, Albrechtslund (2007) takes issue with VSD's positioning as "a principled *and comprehensive* account of human values in design" [italics ours]. Albrechtslund (2007) argues that no design process can be comprehensive with respect to human values; the multistability of human-technology relationships means that no one can fully account for all possible future uses of the designed technology. He cautions against falling prey to "the positivist problem" of assuming that the use of a technology corresponds to its design and against the hubris of assuming that all possible ethical problems with a technology have been accounted for in its design. At the same time, Albrechtslund (2007) acknowledges that many ethical problems can be anticipated and, indeed, that designers have a special obligation to pay attention to unintended uses. Methods developed in the VSD research lab at the University of Washington – notably, value scenarios and the Envisioning Cards – are steps towards helping designers imagine the multiplicity of unintended uses, users, and contexts of use for a technology, as well as the unintended consequences of intended use. Agreeing with Albrechtlund's cautionary statements, we recommend a dose of humility alongside the use of such methods.

Woelfer and Hendry (2011) draw insightfully on the interactional perspective in their discussion of ubiquitous information systems for urban youth experiencing homelessness. Through a value scenario, Woelfer and Hendry (2011) explore the implications of digitizing a youth service agency flyer. This flyer provides the only comprehensive overview of services, the when and where for the youth agencies. Although it is the most comprehensive document concerning services distributed by the agencies, it is not distributed to the public. The print document is available only to users of services who visit the service agencies. Woelfer and Hendry (2011) observe that creating an open, online version of the service agency flyer would provide opportunities to improve usability and access for homeless young people. Yet, making this flyer publicly available on the Internet could also compromise their safety, as its information would be available not only to intended users but also to abusive parents, pimps, and drug dealers. It could bring greater attention – either helpful or harmful – from neighborhood business and home owners. Finally, the easy availability of the flyer online might reduce opportunities for positive face-to-face interactions between homeless youth and the adults who work at the service agencies. Thus, as Woelfer and Hendry (2011) imagine this new tool in its context

of use, they realize that likely implications would distort their original intention. In comparing the online flyer to the paper flyer, it is clear that the use of the technology is determined neither solely by the designers' intentions nor solely by the values of its users but rather by interactions between the properties of the technology (online versus paper), the stakeholders, and the context of use.

*Attention to direct and indirect stakeholders*. VSD "identifies and takes seriously two classes of stakeholders: direct and indirect. Direct stakeholders refer to parties – individuals or organizations – who interact directly with the computer system or its output. Indirect stakeholders refer to all other parties who are affected by the use of the system" (Friedman et al. 2006a). In designing for values, it is important to consider all those who are significantly affected by a technology, not only the clients or users.

Some recent work focuses on direct stakeholders, those whose needs and values will be supported by new technology. For example, Azencot et al. (2011) develop the GoBraille application to support transit users who are blind, or both deaf and blind, while Woelfer et al. (2012) aim to design mobile applications that support the safety of homeless youths. At the same time, other people are recognized to have significant stakes in the technology because of their interactions with the direct stakeholders. Bus drivers could be helped or hindered in their support of blind and deaf-blind passengers (Azencot et al. 2011); service providers, police officers, and community members each have their own relationships with homeless youths (Woelfer et al. 2012).

These researchers design their empirical investigations to invest more effort in direct stakeholders while still including indirect stakeholders. Interviews often require greater mutual investment between researchers and stakeholders than a survey does, but interviews can also reveal greater qualitative detail. Azencot et al. (2011) apply these methods accordingly, interviewing direct stakeholders and surveying indirect stakeholders. They balance their constraints of time and resources, managing to elicit the perspectives of significantly involved indirect stakeholders while concentrating more time on the direct stakeholders. Woelfer et al. (2012) strike a different balance with respect to time investment. They use semi-structured interviews and value sketches to gain rich insights from both direct and indirect stakeholders while tailoring some interview questions to the different stakeholder roles. However, they interview a greater number of direct stakeholders (19 homeless youth) versus indirect stakeholders (four service providers and two police officers), thus investing more time in direct stakeholders while still benefiting from the nuanced perspectives of indirect stakeholders who are nonetheless very involved (Woelfer et al. 2012).

As noted earlier, the same individual can shift between roles, moving from direct to indirect stakeholder. As an example, Czeskis et al. (2011) consider parents and teens as both direct and indirect stakeholders in mobile phone applications designed to support teens' safety. Parents and teens are direct stakeholders when they use these applications. But friendships between teenagers mean that teens and parents who have *not* chosen to adopt such applications are nonetheless affected by their use. For example, a mobile application that takes photographs to monitor a teen's

unsafe driving may also capture images of passengers. An application that discloses a teen's text messages to their parents will also disclose information about those who send and receive the messages. Thus, parents and teens can also be *indirect* stakeholders depending upon the situation. Czeskis et al. (2010) develop value scenarios about exactly these situations. Taking these dual roles further, their empirical investigations ask parents and teens to reflect on both roles: as direct stakeholders who use the technology and as indirect stakeholders who are inadvertently and perhaps unknowingly involved.

For some technologies, large communities have a significant but indirect stakes in the technology's use. For UrbanSim, an urban planning simulator, stakeholders include all residents of the region (Borning et al. 2005). For the "Tribunal Voices" project, stakeholders include all Rwandans (Nathan et al. 2011). Because these important classes of stakeholders are large and diverse, both projects are concerned with engaging stakeholders who have different points of view. Both projects also aim to provide a path for indirect stakeholders to become direct stakeholders: that is, to provide those who are affected by the use of the technology with opportunities to influence its use or appropriate it for their own purposes. In the case of UrbanSim, an important and contested decision is the choice of indicators, or measures, to attend to in interpreting the simulation results. UrbanSim developers engage community organizations to present groups of indicators relevant to their perspectives and recommend indicators for future development; suggested future work would also enable citizens to comment on indicators (Friedman et al. 2008a). Nathan et al. (2011) worked with partner organizations to support workshops on international justice and sexual violence and also to develop online tools that encourage discourse around clips from the Tribunal Voices video collection. These studies suggest ways to consider representation when there is a large, diverse group of people occupying a particular stakeholder role and to include some of those stakeholders as cocreators and users of the technology.

*Tripartite methodology*. As discussed earlier, VSD "contributes a unique methodology that employs conceptual, empirical, and technical investigations, applied iteratively and integratively" (Friedman et al. 2006a). The tripartite methodology can be interpreted rigidly, as though the investigations are lockstep, discrete moves to be undertaken in a prescribed order (Le Dantec et al. 2009). Yet, when looking through recent VSD scholarship, we found many examples of the methodology being applied flexibly, in response to the particulars of the situation and the researchers' goals.

Although the conceptual, empirical, and technical investigations are all considered important, particular studies may rest more heavily on just one or two. For example, a suite of studies starts with a conceptual investigation of the implications of using a digital camera and display to simulate a window in an interior office (Friedman et al. 2006a). But the bulk of the work consists of empirical investigations using multiple methods to address different values and stakeholder roles. One line of research concerns both short- and long-term impacts on the psychological and physiological well-being of those who work in a "room with a view"

(Friedman et al. 2008b; Kahn et al. 2008). Another concerns reflections on privacy by "the watcher and the watched": both the users of the digital display and those whose images are captured by the video camera (Kahn et al. 2008; Friedman et al. 2008c). These studies do not discuss technical investigations informing a product under design; rather, they are looking farther ahead, attempting to understand value implications of hypothetical, near-future technologies. As another example, van Wynsberghe (2013) focuses primarily on a conceptual investigation, applying care ethics to the nascent domain of healthcare robotics in order to identify stakeholders and values at stake. As a brief case study for her design framework, she conducts a technical analysis in which she compares the value implications of alternative designs for robots that assist with lifting patients: an autonomous robot versus a human-operated exoskeleton. van Wynsberghe (2013) indicates that future work applying her framework to design will need to iterate between technical and empirical investigations.

Two recent case studies are particularly instructive in that they discuss the interplay between all three types of investigations within a relatively self-contained design process. The CodeCOOP case study (Miller et al. 2007) presents a full design cycle of software developed with industry partners. In summarizing the design process, Miller et al. (2007) list all major design events and categorize them as conceptual, technical, or empirical. Similarly, Azencot et al. (2011) are careful to articulate the investigations used in their design of the GoBraille tool to support blind and deaf-blind public transit users. Both these design case studies begin with conceptual investigations of stakeholders and values. The CodeCOOP case proceeds to a technical investigation resulting in a software prototype; further technical investigations are interleaved with empirical investigations both formative (surveys, Value Dams and Flows analyses, contests) and summative (usage data analysis, interviews, reflection). One activity – the final design reflection – is considered simultaneously a conceptual and an empirical investigation (Miller et al. 2007). By contrast, the GoBraille case follows the initial conceptual investigation with an empirical investigation: semi-structured interviews with blind transit users, deaf-blind transit users, and an organization and mobility instructor. Building on the stakeholder analysis in the initial conceptual investigation, Azencot et al. (2011) also surveyed bus drivers about their attitudes towards blind and deaf-blind passengers. In technical investigations, Azencot et al. (2011) analyzed existing technologies and found them wanting, developed the low-cost MoBraille platform, and finally built the GoBraille application to support blind transit users in finding stops and identifying their buses. As in the CodeCOOP case, empirical investigations – here, field studies and semi-structured interviews – served to evaluate the new technology. Because the researchers realized their understanding of deaf-blind people was limited, they also codesigned a version of GoBraille with a deaf-blind transit user, thus combining empirical and technical investigations (Azencot et al. 2011). In both these cases, we see an initial conceptual investigation driving an iterative development process interleaving technical and empirical investigations, culminating in an empirical evaluation.

Walton and DeRenzi's (2009) case study of supporting healthcare in Africa is particularly interesting in that the authors do not explicate the integrative nature of their investigations. The case study describes two related but independent projects carried out by each of the coauthors. One project concerns the redesign of existing information technology support for vaccine delivery. Walton and DeRenzi (2009) report that this project engaged with VSD from the beginning; however, the software to be redesigned already existed. In the other project, concerning a tool to support community healthcare workers, VSD is applied to evaluate a proposed design before beginning implementation and user training. Thus, although the application of VSD begins in both projects with a joint conceptual investigation of stakeholders and values, each project is at a different stage of an overall design process when that conceptual investigation was performed. The authors report that the vaccine delivery project continues with a technical investigation – the redesign of the software. However, this technical investigation involves discussion and codesign with stakeholders concerning the meanings of respect and accountability in the context of use, thus taking on an empirical overtone. Walton and DeRenzi (2009) frame their work on support for community health workers as an empirical investigation, including "rapport building," semi-structured interviews, and focus groups with a range of stakeholders. However, this empirical work overlaps with beginning to design and develop the CommCare tool: an iterative process engaging both technical and less formal empirical investigations. This work thus illustrates how investigations overlap and intertwine so that boundaries between them are blurred.

We disagree somewhat with Borning and Muller's (2012) claim that VSD can begin with any type of investigation. Friedman et al. (2006a) recommend a stakeholder analysis as one of the first steps. It does seem difficult to conduct empirical investigations without a reason to engage particular people or to conduct technical investigations with no notion of the user or others who might be affected. Indeed, all three of the case studies discussed above begin their application of VSD with a stakeholder analysis. While Le Dantec et al. (2009) claim that their work begins with empirical investigation, Borning and Muller (2012) point out that the project truly begins with a conceptual move, the identification of homeless people as a stakeholder group worthy of interest. However, we agree with Borning and Muller (2012) that the first *major* investigation can take any of the three forms: for example, a careful conceptual analysis of values at stake, as in the Cookies and Informed Consent work (Millet et al. 2001); an empirical investigation focusing on values in relation to a technology, as in the Watcher and the Watched (Friedman et al. 2006b); or construction of a new technology, as in the CodeCOOP work (Miller et al. 2007). The separation of the investigations is a conceptual tool, a way to get designers to consider the interactional aspects of their work; it is not meant to create silos within the project.

In an attempt to counter the misperception that VSD investigations must proceed in a prescribed order, some recent VSD work focuses on components drawn from the three investigations – for example, indirect stakeholders (conceptual), iterative

practices (technical, empirical), multiple methods (empirical, conceptual, technical), and feature analysis (technical) – rather than calling out the investigations by name. For example, Czeskis et al. (2010) do not refer to the tripartite methodology, but do include conceptual, empirical, and technical investigations in the forms of value scenarios, semi-structured interviews, and technical recommendations. We refer readers to Friedman et al. (2006a) for a discussion of the "Cookies and Informed Consent" case study, which emphasizes the iterative and integrative nature of the investigations using the "traditional" labels. Whether called out explicitly or not, we believe the construct of the tripartite investigation is useful in drawing attention to different ways of exploring the relationship between human values and technology design.

## Heuristics

Beyond VSD's core commitments, we wish to propose several guiding questions based on VSD critiques, VSD case studies, and work that shares an affinity with VSD. We phrase these heuristics as questions addressed to you, the VSD investigator. Note that these heuristics are not strictly orthogonal nor are they presented in the order that they ought to be considered; rather, they intertwine, and your approach to addressing some heuristics will likely affect others. We look forward to future scholarship that continues to develop and add to this list.

*Should you adopt an ethical theory alongside VSD*? As discussed earlier, Friedman's presentations of VSD neither recommend nor forbid the adoption of an ethical theory alongside VSD. An ethical theory can help to identify, define, and prioritize relevant values. Some domains, such as healthcare (van Wynsberghe 2013), war (Cummings 2006), and politics (Borning et al. 2005), have well-developed ethical theories that technology design should draw upon. Although we acknowledge Manders-Huits' (2011) critiques and note that Yetim (2011) proposes discourse ethics as a generally applicable theory, we do not now take a stand as to whether VSD must always be complemented with an ethical theory.

*How will you identify values of concern*? Although the tripartite methodology facilitates the discovery of implicated values throughout the design process, we agree with Le Dantec et al. (2009) that it matters where researchers begin. When design begins with a technology or context of use, rather than a particular value, discovering stakeholders' values through an initial empirical investigation can help to avert researcher bias. However, it can make sense to begin with conceptual investigation when time is short, and especially when the work can build on previous investigations of relevant values (Borning and Muller 2012). Moreover, applying a domain-specific ethical theory can provide a principled basis for identifying relevant values through conceptual investigations.

*Where do you stand on universal values? Does it matter for this project*? We agree with Borning and Muller (2012) that VSD can accommodate different stances regarding the universality of human values. If the researchers' stance on universality

affects their work, they should articulate that stance in reporting the research. The researchers' stance is less likely to be problematic when designers have a nuanced understanding of the situations where their products will be engaged. The researchers' stance is more likely to be an issue when designing across cultures (e.g., as articulated by Alsheikh et al. 2011), when explicitly designing for global use (as noted by Friedman and Kahn 2003), or when anticipating uses far beyond the intended context of use (e.g., with the Envisioning Cards).

*What values*, *if any*, *will the project explicitly support*? In work on UrbanSim, Borning et al. (2005) distinguish explicitly supported values – that is, values which the designers seek to explicitly support during the design process and in the final product – from designers' personal values and from stakeholders' values. Because of its role in a political process, UrbanSim's use engages the full diversity of stakeholder values; the designers' personal values should not be privileged over those of other stakeholders. As Alsheikh et al. (2011) point out, naming values to explicitly support helps prevent the designers' values from being supported by default. To foster legitimacy in the political process, UrbanSim's designers chose three values to explicitly support: fairness, accountability, and democracy (Borning et al. 2005). Borning and Muller (2012) recommend that designers consider participation and power-sharing as explicitly supported values.

*How will you convey stakeholders' voices*? As the previous question makes clear, there are opportunities for researchers to engage with and *hear* stakeholders' perspectives and voice throughout a project. From the initial framing of the work and contexts to the development of theoretical lens, methods, and analysis, opportunities can be created to engage meaningfully with stakeholders. However, how the research team shares the voice of stakeholders with the audience of their work is another question entirely. Space constraints, stylistic norms, and disciplinary conventions can all push against the goal of directly representing stakeholders' voice. However, a community can shift its conventions, particularly when there are strong examples of alternative practices (e.g., Alsheikh et al. 2011; Le Dantec 2009; Woelfer and Hendry 2012).

*How will you present your own voice as a researcher*? Other authors would have presented this overview of VSD in different ways than we have chosen to. Others would have interpreted the VSD literature differently, chosen different points to emphasize, and left different things unsaid. In short, it matters who wrote this chapter. So, too, does the researcher matter in research on design for values. It matters how stakeholders' voices are interpreted, what is emphasized, and what is left out. Ames et al. (2011) and Woelfer et al. (2011) provide particularly strong examples of representing the researcher's voice in design for values. We hope that this chapter serves as an example as well. Throughout this chapter, we have attempted to make clear our voices as authors with particular backgrounds, training, and interests. As we mentioned in the introduction, we have a long history with the VSD lab at the University of Washington. Many of the articles we critiqued have one or the other of us as an author. We believe that making this information apparent is important for positioning you, the reader, in evaluating our claims and entering the discussion.

## Conclusion

Through this chapter, we demonstrate the evolving nature of VSD through the body of scholarship that both critiques and contributes to it since 2006. We conclude with our response to a provocative line of questioning posed by Chris Le Dantec at CHI 2012, in reaction to Borning and Muller's presentation of their paper, "Next Steps for Value Sensitive Design" (Borning and Muller 2012). Le Dantec asked, is it necessary or desirable to continue building a branded approach for engaging with values in the design process? Should not all designers be routinely thinking about values and the human condition and all design educators teaching their students to do so?

We agree with the vision that Le Dantec proposes. But as a field, we are still far from that ideal. Not only is design for human values not yet routine, the development of methods and theories to support that work is still at a nascent stage. Whether the label is VSD or something else, it helps to have a label so that researchers can identify their work in relationship to *something* and build a discourse (as we are doing here) around what the *something* means and how to carry it out.

Consider an analogy with user-centered design, a familiar term within human-computer interaction. According to Abras et al. (2004), the term "user-centered design," or UCD, was coined by Don Norman in the 1980s. This term originally referred to a stance ("the user should be at the center of design"), a theory, and a set of principles. This stance was not obvious at the time, though it became so in retrospect. UCD grew over time, becoming a more general term and less associated with Norman's work, although Norman is still recognized as a founder of the field (Abras et al. 2004).

In the HCI community (broadly conceived), we now take for granted that it is important to consider the user in design and include users in the design process. And yet the term "user-centered design" is still useful, because it distinguishes a user-centered stance from other stances, allowing others to distinguish their work. Moreover, more than 20 years later, software development processes still do not always include attention to the user[6]; that ideal has still not been reached.

A hopeful position is that VSD is the next UCD: Work in the area will continue to grow more nuanced and more reflective. Focusing design on human values will become an accepted rather than novel perspective. Attention to the user is infused throughout HCI work and gaining ground in software development practice, even when there is no explicit reference to UCD; we hope that someday attention to values will be just as pervasive, even if VSD (or another branded values-oriented methodology) is rarely referred to. What we learn from engaging with VSD today will influence how technology designers appreciate and address values in the future.

[6]See, for example, Steve Krug's books *Don't Make Me Think* (2005) and *Rocket Surgery Made Easy* (2010), which introduce software developers to usability principles and practices.

**Acknowledgments** We thank Batya Friedman for suggesting that we write an overview of value sensitive design for this volume. We also thank Batya, along with Alan Borning, Nathan Freier, Peter Kahn, Shaun Kane, and many former lab-mates, for invigorating discussions of VSD. Finally, we wish to thank all those we cite – and particularly those who have so thoughtfully criticized VSD – for joining the discussion and advancing the state of research on design for human values.

## References

Abras C, Maloney-Krichmar D, Preece J (2004) User-centered design. In: Bainbridge WS (ed) Encyclopedia of human-computer interaction. Berkshire, Great Barrington, pp 445–456

Albrechtslund A (2007) Ethics and technology design. Eth Inf Technol 1(9):63–72

Alsheikh T, Rode J, Lindley S (2011) Whose value-sensitive design? A study of long-distance relationships in an Arabic cultural context. In: Proceedings of the ACM 2011 conference on computer supported cooperative work. ACM, New York, pp 75–84

Ames M, Go J, Kaye J, Spasojevic M (2011) Understanding technology choices and values through social class. In: Proceedings of the ACM 2011 conference on computer supported cooperative work. ACM, New York, pp 55–64

Azencot S, Prasain S, Borning A, Fortuna E, Ladner R, Wobbrock J (2011) Enhancing independence and safety for blind and deaf-blind public transit riders. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 3247–3256

Bardzell S (2010) Feminist HCI: taking stock and outlining an agenda for design. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 1301–1310

Barkhuus L (2012) The mismeasurement of privacy. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 367–376

Blevis E (2007) Sustainable interaction design: invention and disposal, renewal and reuse. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 503–512

Borning A, Muller M (2012) Next steps for value sensitive design. In: Proceedings of the 2012 ACM annual conference on human factors in computing systems. ACM, New York, pp 1125–1134

Borning A, Friedman B, Davis J, Lin P (2005) Informing public deliberation: values sensitive design of indicators for large-scale urban simulation. In: Proceedings of the ECSCW 2005 European conference on computer-supported cooperative work. Springer, Amsterdam, pp 449–468

Brey P (2010) Values in technology and disclosive computer ethics. In: Floridi L (ed) The Cambridge handbook of information and computer ethics. Cambridge University Press, Cambridge/New York, pp 41–58

Brodie M, Lai J, Lechner J, Luken W, Ranganathan K, Tang JM (2007) Support services: persuading employees and customers to do what is in the community's best interest. In: Proceedings of the 2nd international conference on persuasive technology Springer, Heidelberg, pp 121–124

Chango M (2007) Challenges to the e-government in Africa south of Sahara: a critical view, and provisional notes for a research agenda. In: Proceedings of the 1st international conference on theory and practice of electronic governance. ACM, New York, pp 384–393

Chatterjee S, Sarker S, Fuller M (2009) A deontological approach to designing ethical collaboration. J Assoc Inf Syst 10(3)

Cockton G (2009a) Getting there: six meta-principles and interaction design. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 2223–2232

Cockton G (2009b) When and why feelings and impressions matter in interaction design [keynote address]. In: Proceedings Kansei 2009: user interface design in practice [Interfejs Uzytkownika-Kansei w praktyce]. Retrieved from http://www.cs.tut.fi/ihte/projects/suxes/pdf/Cockton_Kansei%202009%20Keynote.pdf

Cotler J, Rizzo J (2010) Designing value sensitive social networks for the future. J Comput Sci Coll 25(6):40–46

Cummings M (2006) Integrating ethics in design through the value-sensitive design approach. Engineering and Science Ethics 12(4):701–715

Czeskis A, Dermindjieva I, Yapit H, Borning A, Friedman B, Gill B (2011) Parenting from the pocket: value tensions and technical directions for secure and private parent-teen mobile safety. In: Proceedings of the sixth symposium on usable privacy and security. ACM, New York, p 15

Davis J (2009) Design methods for ethical persuasive computing. In: Proceedings of the 4th international conference on persuasive technology. ACM, New York, p 6

Detweiler C, Hindriks K, Jonker C (2011) Principles for value-sensitive agent-oriented software engineering. In: Agent-oriented software engineering XI. Springer, Heidelberg, pp 1–16

Fallman D (2011) The new good: exploring the potential of philosophy of technology to contribute to human-computer interaction. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 1051–1060

Flanagan M, Nissenbaum H (2007) A game design methodology to incorporate social activist themes. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 181–190

Flanagan M, Howe D, Nissenbaum H (2005) Values at play. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 751–760

Flanagan M, Howe D, Nissenbaum H (2008) Embodying values in technology: theory and practice. In: Van Den Hoven J, Weckert J (eds) Information technology and moral philosophy. Cambridge University Press, Cambridge, pp 322–353

Fleischmann K, Wallace W, Grimes J (2011) How values can reduce conflicts in the design process: result from a multi-site mixed-method field study. In: Proceedings of the American society for information science and technology, vol 48, no 1. ASIS, Medford, New Jersey, pp 1–10

Foong P (2008) Designing technology for sensitive contexts. In: Proceedings of the 20th Australasian conference on computer-human interaction designing for habitus and habitat. ACM, New York, pp 172–179

Friedman B (1999) Value-sensitive design: a research agenda for information technology. Report for value sensitive design workshop. National Science Foundation, Arlington

Friedman B (2004) Value sensitive design. In: Bainbridge WS (ed) Encyclopedia of human-computer interaction. Berkshire, Great Barrington, pp 769–774

Friedman B (2012) Something of value [social impact award talk]. In: SIGCHI conference on human factors in computing systems, Austin, 9 May 2012

Friedman B, Freier N (2005) Value sensitive design. In: Erdelez S, Fisher K, McKechnie L (eds) Theories of information behavior: a researcher's guide. Information Today, Medford, pp 368–380

Friedman B, Hendry D (2012) The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In: Proceedings of the 2012 ACM annual conference on human factors in computing systems. ACM, New York, pp 1145–1148

Friedman B, Kahn PH Jr (2003) Human values, ethics, and design. In: Sears A, Jacko JA (eds) The human-computer interaction handbook. Lawrence Erlbaum, Mahwah, pp 1177–1201

Friedman B, Nissenbaum H (1996) Bias in computer systems. In: ACM transactions on information systems. Conference companion to CHI conference on human factors in computing systems, vol 14, no 3. ACM, New York, pp 330–347

Friedman B, Nissenbaum H (1997) Software agents and user autonomy. In: Proceedings of the first international conference on autonomous agents. ACM, New York, pp 466–469

Friedman B, Kahn PH Jr, Hagman J (2003) Hardware companions? What online AIBO discussion forums reveal about the human-robotic relationship. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 273–280

Friedman B, Freier N, Kahn PH Jr (2004) Office window of the future? Two case studies of an augmented window. In: Conference on human factors in computing systems: CHI'04 extended abstracts on human factors in computing systems, vol 24, no 29. ACM, pp 1559–1559

Friedman B, Kahn PH Jr, Borning A (2006a) Value sensitive design and information systems. In: Zhang P, Galletta D (eds) Human-computer interaction in management information systems: foundations. M.E. Sharpe, Armonk, pp 348–372

Friedman B, Kahn PH Jr, Hagman J, Sevenson R, Gill B (2006b) The watcher and the watched: social judgments about privacy in a public place. Hum Comput Interact 21(2):235–272

Friedman B, Smith IE, Kahn PH Jr, Consolvo S, Selawski J (2006) Development of a privacy addendum for open source licenses: value sensitive design in industry. In: Proceedings of Ubicomp 2006. Springer, Heidelberg, pp 194–211

Friedman B, Borning A, Davis J, Gill B, Kahn P, Kriplean T (2008a) Laying the foundations for public participation and value advocacy: interaction design for a large scale urban simulation. In: Proceedings of the 2008 international conference on digital government research. Digital Government Society of North America, pp 305–314

Friedman B, Freier N, Kahn PH Jr, Lin P, Sodemen R (2008b) Office window of the future? Field-based analyses of a new use of a large display. Int J Hum Comput Stud 66(6):452–465

Friedman B, Höök K, Gill B, Eidmar L, Prien SC, Severson R (2008b) Personlig integritet: a comparative study of perceptions of privacy in public places in Sweden and the United States. In: Proceedings of the 5th Nordic conference on human-computer interaction: building bridges. ACM, New York, pp 142–151

Halloran J, Hornecker E, Stringer M, Harris E, Fitzpatrick G (2009) The value of values: resourcing co-design of ubiquitous computing. Int J CoCreat Des Arts 5(4):245–273

Hofstede G (1991) Cultures and organizations: software of the mind. McGraw-Hill, New York

Iverson O, Halskov K, Leong TW (2010) Rekindling values in participatory design. In: Proceedings of the 11th biennial participatory design conference. ACM, New York, pp 91–100

Johri A, Nair S (2011) The role of design values in information system development for human development for human benefit. Info Technol People 24(3):281–302

Kahn PH Jr, Friedman B, Alexander I, Freier N, Collett S (2005) The distant gardener: what conversations in the telegarden reveal about human-telerobotic interaction. In: Proceedings of 14th IEEE international workshop on robot and human interactive communication, ROMAN 2005. Washington DC, IEEE, Washington, pp 13–18

Kahn PH Jr, Friedman B, Gill B, Hagman J, Severson R, Freier N, Feldman E, Carrere S, Stolyar A (2008) A plasma display window? The shifting baseline problem in a technologically-mediated natural word. J Environ Psychol 28(2):192–199

Kaptein M, Eckles D, Davis J (2011) Envisioning persuasion profiles: challenges for public policy and ethical practice. Interactions 18(5):66–69

Klasnja P, Consolvo S, McDonalid D, Landay J, Pratt W (2009) Using mobile & personal sensing technologies to support health behavior change in everyday life: lessons learned. In: American medical informatics association annual symposium proceedings. MD, AMIA, Bethesda, p 338

Knobel C, Bowker G (2011) Values in design. In: Communications of the ACM, vol 54, no 7. ACM, New York, pp 26–28

Kujala S, Väänänen-Vainio-Mattila K (2008) Value of information systems and products: understanding the users' perspective and value. J Inf Technol Theory Appl 9(4):23–39

Le Dantec C (2009) Situated design: toward an understanding of design through social creation and cultural cognition. In: Proceedings of the 7th ACM SIGCHI conference on creativity & cognition. ACM, New York, pp 69–78

Le Dantec C, Poole E, Wyche S (2009) Values as lived experience: evolving value sensitive design in support of value discovery. In: Proceedings of the 27th SIGCHI conference on human factors in computing systems. ACM, New York, pp 1141–1150

Leitner M, Wöckl B, Subasi Ö, Tschelgi M (2010) Towards the use of negative effects in technology design and evaluation. In: Proceedings of the 24th BCS interaction specialist group conference. British Computer Society, UK, Swindon, pp 443–447

Manders-Huits N (2011) What values in design? The challenge of incorporating moral values into design. Sci Eng Eth 17(2):271–287

Miller JK, Friedman B, Jancke G (2007) Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system. In: Proceedings of the 2007 international ACM conference on supporting group work. ACM, New York, pp 281–290

Millet L, Friedman B, Felten E (2001) Cookies and web browser design. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 46–52

Nathan L, Klasjna P, Friedman B (2007) Value scenarios: a technique for envisioning systemic effects of new technologies. In: CHI'07 extended abstracts on human factors in computing systems. ACM, New York, pp 2585–2590

Nathan L, Friedman B, Klasjna P, Kane S, Miller J (2008) Envisioning systemic effects on persons and society throughout interactive system design. In: Proceedings of the 7th ACM conference on designing interactive systems. ACM, New York, pp 1–10

Nathan L, Friedman B, Hendry D (2009) Information system design as catalyst: human action and environmental sustainability. Interactions 16(4):6–11

Nathan LP (2009) "Ecovillages, sustainability and information tools: an ethnography of values, adaptation, and tension." University of Washington

Nathan L, Lake M, Grey NC, Nilsen T, Utter R, Utter E, Ring M, Kahn Z, Friedman B (2011). Multi-lifespan information system design: investigating a new design approach in Rwanda. In: Proceedings of the 2011 international conference. ACM, New York, pp 591–597

Orlikowski WJ (1992) The duality of technology: rethinking the concept of technology in organizations. Organ Sci 3(3):398–427

Palen L (1999) Social, individual and technological issues for groupware calendar systems. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 17–24

Palen L, Dourish P (2003). Unpacking privacy for a networked world. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 129–136

Pommeranz A, Detweiler C, Wiggers P, Jonker C (2011) Elicitation of situated values: need for tools to help stakeholders and designers to reflect and communicate. Eth Inf Technol 14(4):285–303

Rode JA (2011) A theoretical agenda for feminist HCI. Interact Comput 23(5):393–400

Saab D (2008) An ethnorelative framework for information systems design. In: Proceedings of the fourteenth Americas conference on information systems. Retrieved from http://www.academia.edu/182175/An_Ethnorelative_Framework_for_Information_System_Design

Schwartz S, Bilsky W (1990) Toward a theory of the universal content and structure of values: extensions and cross-cultural replications. J Personal Soc Psychol 58(5):878

Simbasavian N, Jackson M (2007) Designing pervasive brain-computer interfaces. In: Proceedings of the 3rd human-computer interaction and usability engineering of the Austrian computer society conference on HCI and usability for medicine and health care. Springer, Heidelberg, pp 267–272

Singley K, Lai J, Kuang L, Tang JM (2008) BlueReach: harnessing synchronous chat to support expertise sharing in a large organization. In: CHI'08 extended abstracts on human factors in computing systems. ACM, New York, pp 2001–2008

Timmermans J, Zhao Y, van den Hoven J (2011) Ethics and nanopharmacy: value Sensitive design for new drugs. NanoEthics 5(3):269–283

Ulrich W (2000) Reflective practice in the civil society: the contribution of critically systemic thinking. Reflect Pract 1(2):247–268

Van Wynsberghe A (2013) Designing robots for care: care centered value-sensitive design. Sci Eng Eth 9(2):407–433

Walton R, DeRenzi B (2009) Value-sensitive design and health care in Africa. Prof Commun IEEE Trans 52(4):346–358. IEEE, Washington

Woelfer J, Hendry D (2009) Stabilizing homeless young people with information and place. J Am Soc Inf Sci Technol 60(11):2300–2312

Woelfer J, Hendry D (2011) Designing ubiquitous information systems for a community of homeless young people: precaution and a way forward. Pers Ubiquitous Comput 15(6):565–573

Woelfer J, Hendry D (2012) Homeless young people on social network sites. In: Proceedings of the 2012 ACM annual conference on human factors in computing systems. ACM, New York, pp 2825–2834

Woelfer J, Iverson A, Hendry D, Friedman B, Gill B (2012) Improving the safety of homeless young people with mobile phones: values, form and function. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 1707–1716

Woelfer J, Yeung M, Erdmann C, Hendry D (2008) Value considerations in an information ecology: printed materials, service agencies, and homeless young people. In: Proceedings of 71st annual meeting of the American society for information science and technology, vol 45, no 1. ASIS, Medford, New Jersey, pp 1–9

Woelfer J, Iversen A, Hendry D, Friedman B, Gill B (2011) Improving the safety of homeless young people with mobile phones: values, form and function. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 1707–1716

Wyche S, Taylor A, Kaye J (2007) Pottering: a design-oriented investigation. In: CHI'07 extended abstracts on human factors in computing systems. ACM, New York, pp 1893–1898

Wyeth P (2006) Ethnography in the kindergarten: examining children's play experiences. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 1225–1228

Yetim F (2008) Critical examination of information: a discursive approach and its implementations. Informing Science: the International Journal of an Emerging Transdiscipline 11:125–146.

Yetim F (2011) Bringing discourse ethics to value sensitive design: pathways to toward a deliberative future. AIS Trans Hum Comp Interact 3(2):133–155. ACM, New York,

# Participatory Design and Design for Values

Maja van der Velden and Christina Mörtberg

## Contents



**Abstract**

Participatory Design (PD) is a design methodology in which the future users of a design participate as co-designers in the design process. It is a value-centered design approach because of its commitment to the democratic and collective

M. van der Velden (✉)
Department of Informatics, University of Oslo, Oslo, Norway
e-mail: majava@ifi.uio.no

C. Mörtberg
Department of Informatics, Linnaeus University, Växjö, Sweden
e-mail: christina.mortberg@lnu.se



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_33

shaping of a better future. This chapter builds forth on the Scandinavian Participatory Design tradition. We discuss why the design process is as important as the final result, the product, or service. The creative application of Participatory Design methods facilitates a design process in which values emerge and become inscribed in a prototype. We present PD's guiding principles: equalizing power relations, democratic practices, situation-based action, mutual learning, tools and techniques, and alternative visions about technology. In addition, we discuss some value practices and design methods informed by our PD projects in health care and the public sector. We maintain that Participatory Design increases the chance that the final result of a design process represents the values of the future users.

**Keywords**

Participatory Design guiding principles • Emergent values • Participatory Design methods

## Introduction

Participatory Design (PD)[1] is a collection of design practices for involving the future users of the design as co-designers in the design process. PD's methodology is based on the genuine decision-making power of the co-designers and the incorporation of their values in the design process and its outcome, which is often a high-fidelity prototype for a product or service, or a new way to organize a work practice or to design a space. The core theme of Participatory Design, as formulated by Pelle Ehn (1988), is addressing the dialectics of tradition and transcendence: the tension between *what is* and *what could be*. PD's methods enable participants to anticipate future use and alternative futures.

In the most recent comprehensive volume of writings about Participatory Design (Simonsen and Robertson 2012), PD is defined as follows:

> A process of investigating, understanding, reflecting upon, establishing, developing, and supporting mutual learning between multiple participants in collective 'reflection-in-action'. The participants typically undertake the two principal roles of users and designers where the designers strive to learn the realities of the users' situation while the users strive to articulate their desired aims and learn appropriate technological means to obtain them. (p. 2)

The Participatory Design tradition was established in Scandinavia in the early 1970s. It was influenced by, and developed concurrently, with a range of projects

[1]We use the term Scandinavian Participatory Design (PD) to refer to the early years of the Participatory Design tradition in the Scandinavian countries and Participatory Design (PD) to refer to the design tradition in general. Several PD researchers cited in this chapter use or have used different terms to refer to the early years of the Participatory Design tradition, such as user-centered systems design, systems development, and cooperative design.

with a focus on the democratization of the working life. These projects were conducted by the trade unions or jointly by the trade unions and working life researchers. The Norwegian Union of Iron and Metalworkers (NJW) initiated one of the first PD projects in cooperation with researchers from the Norwegian Computing Center from 1970 to 1973. The objective was to involve the workers in the design of a computer-based planning and control system for their workplace. A plan was designed, based on a participative approach and the inclusion of workers' knowledge, with several activities for the unions, including working groups to discuss and to find solutions through action programs, assessments of existing information systems, and propositions of changes. The researchers participated with lectures as well as with support in the development of the project results (Nygaard and Bergo 1975). In addition, educational material, in the form of a textbook, was developed to support these activities (Nygaard 1974).

Similar projects followed in other sectors, such as the Swedish project *Democratic Control and Planning in Working Life* (DEMOS)[2] (Ehn and Sandberg 1979) and *Office Automation* (Morgall and Vedel 1985). The Scandinavian PD tradition did not only establish new ways to involve workers in IT design projects, it also helped establish new Data Protection Acts and influenced the Worker Protection and Working Environment Acts in Norway, Denmark, and Sweden. In addition, these projects brought also the psychosocial issues of the working life to the foreground (Bratteteig 2004), and new notions were invented to adapt the specialist language of the researchers to the local and situated practices (Nygaard and Bergo 1975).

Having a Say, the title of Kensing and Greenbaums's (2012) chapter in *Routledge International Handbook of Participatory Design* (Simonsen and Robertson 2012), reflects the main focus in the early days of PD. It was essential to engage those who were going to be affected by the development and implementation of an IT system in decisions about the design of the system: workers were having decision-making power in the design of new technologies that would affect their work and their skills.

Kensing and Greenbaum consider political economy, democracy, and feminism as the theoretical roots of the early Scandinavian PD tradition. Participatory Design evolved in a time in which workers and activists organized themselves to demand improved working conditions. Their organizations challenged the asymmetrical power relations between trade unions and employer organizations/management and demanded the inclusion of the voices of those in the margins in order to increase their influence at the societal level. The third root of Participatory Design, feminism, helped focus on work dominated by women and to include their voices and skills in the design process. For example, the Florence project (1984–1987) at the University of Oslo involved nurses – a professional group dominated by women – in the design of an IT system (Bratteteig 2004).

[2] 1975–1980

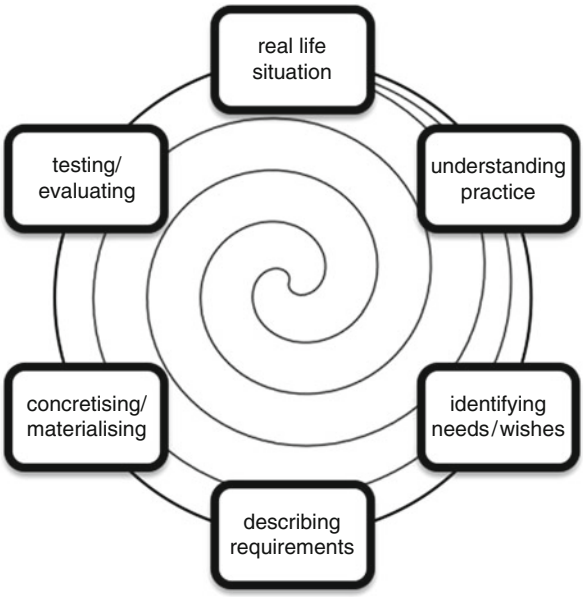



**Fig. 1** Use-oriented design cycle (Based on Bratteteig et al. 2012, p. 128)

Bratteteig et al. (2012) describe several design approaches found in Participatory Design. Our own PD practices, in which we focus on the design of technologies and services, are mainly informed by the use-oriented design approach, which is based on a six-phase iterative design cycle (see Fig. 1). In this approach, process and product are of equal importance. The design process enables the emergence of values and definitions of use, while the artifact (product or service), in its different stages of development, enables the exploration of those different definitions of use (Redström 2008).

As we will discuss below, the values emerging during this design process are materialized in the designed object. The designers and co-designers take design decisions that implicitly and explicitly inscribe values in the final product. Together they present the design's meaning and prescription or *script* for use (Verbeek 2006). The importance of this process lies in the fact that technology mediates the behavior of people. The script of a designed product promotes certain use(s) and inhibits other use(s) (ibid.). When there is a discrepancy between the design context and use context, the script will not be strong or stable enough to mediate the behavior of the users in the way it was envisioned (Latour 1991). Participatory Design can offer a democratic way to "moralize technology" (Verbeek 2011). At the same time, we agree with Ihde (1999) and Albrechtslund (2007) that technology is multi-stable, that is, it can have different stable meanings in different use contexts. Participatory Design does not promise a direct connection between design and use, somehow forcing a design's meaning on all use contexts. On the other hand, involving future users as co-designers in the design process significantly increases the chance that the product represents the values and meaning of the future users.

## Participatory Design's Guiding Principles

Participation and democracy are the principal values of PD (Bratteteig et al. 2012; Robertson and Wagner 2012). These values challenge traditional systems design approaches, which are based on a distance between designers and prospective users of the projected information technologies. Having a say in the design process, in all its activities and decisions, enables other principles, such as a design practice based on *equalizing power relations* (Kensing and Greenbaum 2012). In addition, the design process involves co-realization with a range of participants with their diversity of experiences and knowledge (Bratteteig et al. 2012). The commitment to *democratic practices* results in involving all those who will be affected by the new technology in the design process (Kensing and Greenbaum 2012). These democratic practices are to be maintained throughout all design activities, enabling trust among all those involved and facilitating a learning process and a commitment to taking responsibility for each other and for the design result. Besides *equalizing power* and *democratic practices*, Kensing and Greenbaum mention four other guiding principles: *situation-based actions*, *mutual learning*, *tools and techniques*, and *alternative visions about technology*.

### Situation-Based Action

Situation-based action pays attention to people's expertise of their day-to-day activities in work or other practices. In doings and actions, individually or collectively with other people and technology, skills and knowledge are shared and gained. Thus, design is always performed *somewhere* by humans and nonhumans; their activities do not take place in isolation but are embodied and situated (Suchman 2002, 2007). People's skills and expertise are implicated in both visible and invisible work (Elovaara et al. 2006; Karasti 2003; Star and Strauss 1999). To locate the design activities in people's daily work or other doings may avoid understanding work only as "organizational processes," and approaches work as being "part of the fabric of meanings within and out of which working practices – our own and others' – are made" (Suchman 1995, p. 58).

### Mutual Learning

Democracy and participation also enable *mutual learning*, learning from each other. Co-designers – such as workers – learn from designers about design and related technological issues. In turn, designers learn about the workplace – the use context – and the workers' activities and skills. In mutual learning, the participants not only share their practical knowledge, competences, and values, "they also learn more about their work themselves" (Karasti 2001, p. 236). Karasti came to this conclusion through Participatory Design workshops held at the radiology department at the Oulu University Hospital, Finland. The workshops provided the

participants the opportunity to scrutinize their own work practices. She continues: "The analytic distance allowed them to articulate meanings of work and to discover previously invisible taken-for-granted aspects of routine practices" (Karasti 2001). Jansson (2007) confirms Karasti's findings: embodied knowledge is implicated in people's day-to-day work and becomes visible in participatory activities.

Although the underlying value, to learn from each other, is implicated in the notion of mutual learning, there is a risk of falling back into dualities and thereby getting caught up in keeping apart the designers and other participating practitioners in the design process. In the dominant discourse in participatory methodologies, there are tendencies to mainly focus on the users and the use context without paying attention to the designers' values and norms and how they are brought into the design setting (Markussen 1996). In addition, the taken-for-granted-ness of professional designers' expertise has also been questioned (see Bath 2009; Vehviläinen 1997). Researchers have used the notion of *design from nowhere* to refer to the disembodied or even invisible professional designer, who is located everywhere (Markussen 1996; Suchman 2002). In *design from somewhere*, Suchman (2002) takes the reverse position: knowledge, including professional expertise, is understood as partial and situated knowledge (Haraway 1988), dependent on a practice, its history, the involved participants' views and knowledge, and their participation.

Today, Participatory Design is a methodology used in several other disciplines, such as urban planning, architecture, and sustainable development. It also moved from local to global settings, while its methodology has also been challenged by changes in societies and the development of new technologies since the early days of PD, such as transformations in the socioeconomic makeup of societies; development of personal technologies; the diffusion of information technology to every aspect of everyday life; and, most importantly, more knowledgeable co-designers. For example, the Health Information Systems Programme is a Participatory Design project developing free and open source District Health Information System (DHIS) software. The project started in South Africa in the early 1990s but has since developed in a global network of participating countries and institutions. The meaning of participation, and the methods to support participation, has changed significantly, from design workshops involving public health researchers and activists with a background in the anti-apartheid struggle and informatics researchers to a global virtual team of developers in a South-South-North network (Braa and Sahay 2012).

Ubiquitous computing, smartphones, and other technologies are now available everywhere, at home, at work, and in public spaces, challenging the dominant expertise discourse, which in turn challenges Participatory Design. For example, when the Scandinavian PD tradition started, the knowledge of information technology was something only technical experts possessed (Markussen 1996). These days, co-designers can be expert users in the technology under design, such tech-savvy teenagers in a social media and mobile applications project (e.g., van der Velden and Machniak 2014) or bioinformaticians in a participatory programming project (Letondal and Mackay 2004).

## Tools and Techniques

One way to practice participation is by using participatory methods, or *tools and techniques*, which forms another important principle in PD (Kensing and Greenbaum 2012). A range of methods have been developed and introduced to facilitate participation and cooperative design, such as future workshops, mock-ups, storyboards, scenarios, probes, walk-throughs, games, workshops, cartographic mappings, collaborative prototyping, etc. (Bødker et al. 2004; Brandt et al. 2012; Bratteteig et al. 2012; Mörtberg et al. 2010). Ethnography is also widely used in PD, particularly in the initial phases, to capture the richness of work and other practices (Blomberg and Karasti 2012). The ethnographer may also work as a facilitator to create dialogues and to enable collaboration (Blomberg et al. 1993). In these kinds of collaborations, everyone contributes with their knowledge and perspectives in participatory dialogues (Christiansen 2014; Luck 2003), with the aim to bridge the distance between the various practitioners and to enable questions about "the terms 'we' and 'others'" (Suchman 1995, p. 61). All PD methods try to encourage participatory dialogues and to integrate ethical values in the entire design process. In section "Participation, Methods, and Values," we will discuss four of such participatory methods.

Methods, or tools and techniques, play a central role in the creation of an inclusive and democratic design space: "A major strength of Participatory Design is that there is a robust connection between ethical practice and the choice of methods, tools, and techniques" (Robertson and Wagner 2012, p. 78). Participative methods are a prerequisite to enable people to participate in the design process as experts of their day-to-day work or daily life. Telier et al. (2011) argue this is central to PD: "[In fact, as we shall see,] the origination of participatory design as a design approach is not primarily designers engaging in use, but people (as collectives) engaging designers in their practice" (p. 162).

## Alternative Visions About Technology

The early PD projects made clear that *having a voice* in a design process was not enough. *Having a say* in the technology design was necessary for real change to take place: "It appeared to be necessary to create alternative technologies as well as to fight vendors' monopoly over technological choices" (Kensing and Greenbaum 2012, p. 29). Developing an alternative vision of technology was, for instance, central to the UTOPIA project, which took place in Sweden from 1981 to 1984. For this purpose, a lab-like design space was created in which workers and researchers could jointly experiment with scenarios, paper mock-ups, and different technological solutions. Developing alternative visions about technology is still a crucial aspect of PD. It often involves a design process in which an existing technological solution is redesigned or replaced with an alternative solution based on the values of the users and not those of the management, vendor, or owner.

The guiding principles of Participatory Design are the result of experiences and practices in the early PD projects, which were politically motivated. The guiding principles give Participatory Design also an ethical orientation, as they focus on participation, inclusion, equality, and sharing. In the following section, we will discuss this orientation with a focus on values.

## A Value-Centered Design Approach

In their seminal paper "User Participation and Democracy: A Discussion of Scandinavian Research on System Development," Bjerknes and Bratteteig (1995) argue that there has been a turn from politics to ethics in research on user participation in Participatory Design. Focusing on Scandinavian projects supporting *workplace* democracy, the authors differentiate between a political and ethical road to democracy. In the 1970s, system development projects had an explicit political character. Their mandate included pursuing change in laws and agreements that regulated the use of computers in the workplace. The role of the system developer was that of an emancipator. This changed, according to the authors, in the second half of the 1980s. The system developer as an emancipator in a collective political process became the system developer as a professional facilitator of a system development process. The ethical responsibility of the system developer is now based on his/her own individual ethics, which might or might not be supportive of a larger political program.

Although Bjerknes and Bratteteig argued in their paper for a reintroduction of the collective political dimension in system development, in order to contribute to workplace democracy and workers' rights, the *ethical road* became the hallmark of Participatory Design approach. Participatory Design was presented as a more ethical approach to the design of information systems and other technologies (Robertson and Wagner 2012). Steen (2013) refers to this shift from politics to ethics as the *ethical turn* in Participatory Design.

### Values and Ethical Motivation

The discussion of whether politics or ethics, or both, hallmark Participatory Design is partly inspired by discussions on the definitions of and relation between politics and ethics. We take a pragmatic perspective and state that Participatory Design is a value-centered approach to design, since both politics and ethics are value-centered theories. That is, both politics and ethics want to realize values, "the idealized qualities or conditions in the world that people find good" (Brey 2010, p. 46).

Secondly, Participatory Design is a value-centered design approach because of its *ethical motivation*, which is built on values. Supporting and increasing democratic practices still is an ethical motivation for Participatory Design (Bjerknes and Bratteteig 1995; Ehn 1988; Robertson and Wagner 2012). The latter defines PD's ethical motivation as "to support and enhance how people can engage with others in

shaping their world, including their workplaces, over time [. . .] working together to shape a better future" (ibid., p. 65). Robertson and Simonsen (2012, p. 5) define this ethical stand of Participatory Design as the recognition of "an accountability of design to the worlds it creates and the lives of those who inhabit them." These descriptions reflect the way in which Participatory Design has broaden its field, moving out from Scandinavian system development focusing on workers' rights and workplace democracy to a more encompassing general and global notion of accountability and *shaping a better future*.

## Emerging and Dynamic Values

Participatory Design projects do not "frontload" (van den Hoven 2007) a fixed set of values, but they may frontload a set of moral values appropriate to the particular context. This can be explained by PD's central focus on participation, which results in an approach in which the design process is as important as the end result of this process, the designed. All participants in the design process *have a say* in which moral values should inform the design process and the designed. In other words, moral values may also emerge and be cocreated during the design process. Participatory Design's tools and techniques support such emergence of values as well as help deal with conflicting values (see Iversen et al. 2012).

Values also can stay implicit or latent in the design process. Halloran et al. (2009) give the example of three cases in which values only emerged after the participants were challenged by others or by particular developments in the design process. Most importantly, they found that the relationship of values to design is dynamic (ibid):

> [V]alues emerge during co-design work whether or not we look for them. In addition, there is value evolution, values can change and even conflict as the design process unfolds. This bottom-up, data-driven approach to value identification can provide leverage in solving a number of practical co-design problems as the process unfolds; as well as focusing design activity relevant to the users, it can help with the alignment of values between researchers and users, supporting the design relationship, helping users to understand and contribute at functional and technical levels, lead to user insight about their own values and enable the expression of values both during the design process and, ultimately, in the designed artefact. (p. 271).

## Whose Values

Value Sensitive Design (VSD) is another well-known approach to design for values. The aim of VSD is to advance moral values in design, in particular human welfare, ownership and property, privacy, freedom from bias, universal usability, trust, autonomy, informed consent, accountability, identity, calmness, and environmental sustainability (Friedman and Kahn 2003). Manders-Huits and Zimmer (2009) see an important difference between VSD and Participatory Design.

They acknowledge Participatory Design as a value-centered approach but argue that PD is "falling short of directly addressing values of moral import, such as privacy or autonomy" (p. 58). They differentiate between design frameworks that "seek to broaden the criteria for judging the quality of technological systems to include the advancement of moral and human values," such as Value Sensitive Design, and design frameworks such as Participatory Design that promote functional or instrumental values (see also Manders-Huits and van den Hoven 2009). Their evaluation raises an important question: *whose values* are advanced in these value-oriented design perspectives?

The question of *whose values* was explored early on in the development of the Participatory Design approach (Wagner 1993, p. 100):

> This raises the larger issue of how "egalitarian values" – equality, inclusivity, sharing, participation – be implemented. [. . .] In question is also the role of systems designers' own values in a PD project. Some may have a tendency to view themselves as impartial deliverers of technical support to organizational "solutions" worked out by users. Others may argue that systems designers' values inevitably enter the process of negotiating conflict over ethical questions.

Values have also become a site of cultural and generational conflicts. In multicultural settings, the principle values of Participatory Design, participation and democracy, need to be explored and understood in the local context (Elovaara et al. 2006; Puri et al. 2004). For example, an ongoing Participatory Design project in Namibia shows that participation is not necessarily associated with democracy (Winschiers-Theophilus et al. 2012): "In a hierarchical society lower ranking members are not expected to publicly and openly express opinions, although they are not formally prohibited from doing so. This might seem unjust and counterproductive to participation, when participation is associated with egalitarianism or democracy" (p. 165). A PD project in Cambodia, designing a device that would enable children using a prosthetic leg to walk in the mud, showed that cultural and socioeconomic structures prevented a participatory process involving all stakeholders: "the users were raised in a culture where children are not encouraged to express their own opinions but to be obedient towards adults" (Hussain et al. 2012). The designers were able to solve the problem by organizing separate design workshops for adults and for children.

## Value Practices in Participatory Design

Values play a central role in Participatory Design. The principal values of participation and democracy are perceived as PD's central values as they inform PD's approach and methods. Methods used during the earlier phases in the design process (see Fig. 1) enable the emergence of the needs and values of the co-designers. In Scandinavian PD projects of the 1970s and 1980s, the values of worker participation and workplace democracy, together with quality of working life, were also considered the central values informing the designed object (Kensing and Greenbaum 2012). Iversen et al. (2012) argue against privileging values, including

those values associated with PD, such as "participation, democracy, and human welfare" (p. 90). Their design approach focuses on an appreciative judgment of values by the designer through a dialogical process of the emergence, development, and grounding of values. This dialogical process is also used to overcome value conflicts. Appropriate methods are brought in to help the different co-designers to reengage with their values: "The idea is to create opportunities for them to question and to renegotiate their values, potentially unfinalising their original perceptions of their values. Sometimes, this could even lead to new conceptualisations of their values" (p. 96).

In our own design practices, we often use a combination of value practices. We facilitate a process in which values can emerge, but we often also need to *frontload* certain values (van den Hoven 2007) when they are part of the design brief of a design project. For example, in our design practices in the health-care sector, autonomy and privacy are central moral values, but we experienced different understandings of these values between young co-designers and ourselves, especially in an online context. In the same practices, autonomy and privacy are also instrumental values, as patient autonomy and personal health information are regulated by laws and regulations and are enforced through Research Ethics Board reviews and informed consent forms. Sometimes these different meanings of the same values can be in conflict. What autonomy and privacy mean, and how they become materialized in a design, emerges in and through the design process. We thus understand the design process as a *contact zone* (Haraway 2003; Pratt 1998; van der Velden 2010), in which different meanings and understandings of autonomy and privacy *meet* and *grapple* with each other. Such meanings and understandings do not meet as wholes; they are relational entities that enter new relations in the design process. The notion of contact zone helps us to understand that the design process is a space for "communication across irreducible differences" and "situated partial connections" (Haraway 2003, p. 49). We may never fully know each other's values, but we can meet *respectfully* in our design activities. The design process thus becomes a space for a pluralism of values and for an *agonistic, noncoercive consensus* (Mouffe 1993).

The dialogical process described by Iversen et al. (2012), which is based on the analysis of three PD projects, is a good example of the PD process as a contact zone. In a dialogical process, the participants do not take nonnegotiable positions. Through discussions, observations, visualizations, and interpretations, participants were able to *meet* and *grapple* with other positions, which resulted in the renegotiation of their own positions. For example, in the case of the design of an interactive school project (Iversen et al. 2012, p. 96), the students' values oriented the project towards a student-centered model, questioning the central role of the teachers. The designers introduced fictional inquiry (Dindler and Iversen 2007), similar to *future workshops* (see section "Participation, Methods, and Values") to facilitate a fictional space in which both students and teachers could play their roles without being threatened. The result was that the teachers began to explore their role as game masters, thus reconceptualizing the value of what it means to be a teacher and educator.

Paying attention to multiple voices is foundational in Participatory Design, but this can result in value conflicts. A conflict can become an important resource in the design process (Gregory 2003) but can also be the result of larger organizational conflicts, which may be *undissolvable* (Bødker 1996). When the value conflict is located in the group of participants itself, PD's wide variety of tools and techniques can be used to explore the different values. Rather than using tools to build consensus, tools and techniques are used to explore this value pluralism while postponing decisions on the formulation of problems and solutions (Bratteteig and Wagner 2014, pp. 19–21). Also the Aristotelian concept of *phronesis* has guided dealing with value conflicts. As Ehn and Badham (2002) write:

> In phronesis, wisdom and artistry as well as art and politics are one. Phronesis concerns the competence to know how to exercise judgement in particular cases. It is oriented towards analysis of values and interest in practice, based on a practical value rationality, which is pragmatic, and context dependent. Phronesis is experience-based ethics oriented towards action. (p. 6)

Flyvbjerg describes *phronetic research* as research with a focus on power and values and the task of the phronetic researcher "to provide concrete examples and detailed narratives of how power works and with what consequences, and to suggest how power might be changed and work with other consequences" (Flyvbjerg 2001, p. 140). Phronesis can be understood as one of the ethical roots of Participatory Design, enabling imagination and emotions to have a part in the design process (Bratteteig and Stolterman 1997) and to guide the process "to serve the common good and avoid harming people's possibilities to develop a life of their own" (Kanstrup and Christiansen 2006, p. 328).

## Participation, Methods, and Values

Methods are central to creating an inclusive and democratic design process: they help define and facilitate participant and participation, they enable the expression and exploration of values, and they make use-before-use possible (as will be explained below). We started this chapter with a brief recollection of Participatory Design of the 1970s, which illustrated how participation and its related values became the principal elements in Participatory Design. "[G]enuine participation" is considered both a political and ethical value and the core of Participatory Design (Bødker et al. 2004). In this context, Robertson and Simonsen speak of the "fundamental transcendence of the user's role from being merely an informant to being a legitimate and acknowledged participant in the design process" (2013, p. 5). But what is *genuine* participation? In *Disentangling participation: Power and decision-making in Participatory Design*, Bratteteig and Wagner (2014) explore what they call the most difficult part of Participatory Design: "the sharing of power inherent in the PD approach: in order to collaborate with users as co-designers, the designers need to share their power with them and acknowledge their different and equally valuable expertise" (p. 6). With the spread of Participatory Design to other

areas and sectors, participation sometimes became an instrumental value, no longer based on sharing decision-making power with participants, and exploitative (Keinonen 2010; Shapiro 2010). Genuine participation requires participatory designers to be mindful about the issue and workings of power.

One important thread in the focus on participation has been the relation of the designers and researchers to the other participants in the design process. In early Participatory Design, they were workers, but their role in the design process was not perceived as part of their position in their workplace. Their role was defined as (future) users of the system under design. This was reflected in the early Scandinavian Participatory Design literature, which uses the term user participation, not worker participation. Greenbaum and Kyng (1991) speak of the *awkwardness* of calling the people, who use computers, users, in the context of a participatory design process, but continue using the term nevertheless. This issue continued to occupy PD researchers. Bødker (1996) wrote: "[. . .] the term user may be a mistake. However, for lack of a term that covers all the different kinds of workers [. . .] I will utilize the term in this article" (p. 217). Redström (2006, 2008) argues that calling people users, while the things are not yet designed, obscures the process of becoming a user. We therefore prefer to use the term *co-designers* for the people with whom and for whom we design.

## Multiple Voices and Silences

The focus in a technology design process is often on what seems the most obvious solution to a problem: to design a technology that can do something more efficient, effective, smarter, or entertaining. In Participatory Design, we try to postpone *the most obvious solution* in order to enable co-designers to explore alternative visions of technology and their different uses and effects. This is facilitated by Participatory Design's commitment to democratic practices, which enables people who otherwise might be invisible or marginalized to have a voice in the design process. Elovaara, Igira, and Mörtberg (2006, p. 113) present the following notes from a research diary of a conversation in a design process:

> We are laughing and talking. One of our participants is telling a wonderful story from her everyday work. One citizen visited her. The woman was behind with the payments of day nursery. She wanted to pay in cash but the system did not accept any real money. So the civil servant said to the person: "Let's go together to the bank. Then I'll see to it that you fix it and I get a receipt. When I go back to the office I register your payment. After that everything is in order". And that's what they did: they marched together to the nearby bank. Everything worked out smoothly. And everybody was happy: the citizen (she could pay the fee for day nursery and not end up in the enforcement register), the civil servant (she could do her job and receive the expected fee) and the municipality (they got their money). Then Pirjo [one of the design researchers] asked if citizens couldn't use the municipality web site to check and pay their taxes and other fees – develop a self-service municipality? We started to imagine and make scenarios. And suddenly I heard the civil servant who sat next to me, she said, very, very quietly: "But then Anna (referring to the civil servant working with the municipal fees, sorting them, posting the payment forms, reminding etc.), will become unemployed?!."

The risk to be replaced by technology was the value that emerged from the participative methods in the project. Elovaara et al. (2006) refer to this as an emergent ethical issue related to knowledge and visibility. In order to *do no harm* to the co-designers, they stress the need to "take the PD core ideas seriously" and to find alternative visions to what was the most obvious solution, that is, to replace a person with technology.

To pay attention to multiple voices is central in PD, including voices that are silent or are silenced. Stuedahl (2004) discusses silences in a Norwegian Participatory Design project called NEMLIG (net- and multimedia-based learning environments – 2000–2002). The project was conducted at a middle-sized Norwegian graphical company, and the purpose was to design net-based features for learning at work. The participants involved in the projects were designers, researchers, graphical workers, typographers, and graphical designers. The pilot was managed by a union-based competence center, which was not well versed in participatory design methods. The competence center organized a range of design meetings. The third design meeting resulted in a breakdown when the graphical workers stopped the design process. Mörtberg and Stuedahl (2005, p. 142) write:

> This is what happened during a workshop session, when the systems developer explained possible solutions for the system: the users turned silent. They did not ask questions, and they did not comment on the ideas the system developers presented. After one hour of silence – meaning users' silence, while the system developers discussed a variety of issues – the head of the user group spoke up and asked the system developers [for a break].

The break had the effect of a breakdown (Luck 2007). Afterwards, the co-designers were able to express their grievances and to bring their values and needs into the design process.

Silences may emerge during a design process and it has implications for the participants' actions. The challenge is how to respect silences and to create a democratic design space, which gives voice to all, included to those who are silent. Sensitivity, as a virtue, is necessary to pay attention to silences (Karasti 2001; Mörtberg and Stuedahl 2005), as there is a tendency to neglect silences (Finken and Stuedahl 2008).

The two examples of our research diaries show that a participatory design process is not only about designing an object. The designers and co-designers work together in the design process, making visible and concretizing their knowledge, values, and needs into the designed. Simultaneously, understandings of the use(s) and users of the design emerge in this iterative process. For example, Anna, in the first example, emerges as a nonuser because she would lose her job if the municipality would start using an online payment system. The co-designers in the second example first emerge as nonusers and express this through their silence. After the breakdown in the process, the co-designers emerge as potential users, depending on how the designers would address their needs. The central design challenge in both cases is that the *use* of the designed object can only be envisioned or anticipated, because there is no actual artifact that can be used: this is what is called *use-before-use* (Ehn 2008; Redström 2006). Use-before-use requires participatory methods that engage the co-designers in telling, making, and enacting use (Brandt et al. 2012).

## Values and "Designing for Use-Before-Use"

Participatory designer Pelle Ehn has described PD as "designing for use-before-use" (2008). Over the years, a range of methods have been designed, developed, and evaluated that engage and support co-designers in envisioning the use of the designed. These methods aim to express and explore values as well as to inscribe these values in the product of the design process, the materialization of values (Verbeek 2006). Although all participatory design methods and activities elicit information, discussion, reflection, and learning, some methods are particularly suited for expressing, exploring, and materializing values by engaging co-designers in telling, making, and enacting use.

Certain PD methods are particularly productive for exploration and engagement in the early stages of the design process, when the focus is on *understanding practices* and *identifying needs and wishes* (see Fig. 1). During a Participatory Design process, values are expressed and explored and then become materialized in the form of an object. The materialization of values is the result of interactions between the designers and co-designers and material objects (materials, tools, mock-ups, prototypes, etc.). In this process, co-designers become users and the material object becomes a product or a service. "Technology is society made durable," wrote Bruno Latour (1991). In a participatory design process, ideas of what is good (values), useful (functions), and beautiful (aesthetics) are made durable in inscriptions in the design, such as choices in materials, options for use, color use, flexibility, etc. Anthropological and sociological studies of the design and use of technology have shown that these inscriptions of values in a design are not prescriptive. Although they can function as a *script* for use (Akrich 1992), if and how an artifact will be used is based on the strength of the inscriptions (Latour 1991). On the other hand, the fact that future users of an artifact have been involved in the inscription process often results in artifacts that matter to them and increases the likelihood that they will use it the way it was envisioned.

## Methods for Exploring, Engaging With, and Materializing Values

There is a large body of design methods in use in Participatory Design (e.g., Brandt et al. 2012; Sanders et al. 2010; Spinuzzi 2005; Wölfel and Merritt 2013). We will discuss four of these methods: card methods, mapping methods, future workshop, and participatory prototyping (see Table 1).

### Card Methods

Cards and card sorting methods are widely used in PD. Cards are rectangular or square pieces of paper or carton (approximately the size of playing cards) containing text, an image, or both. They have a large use potential because of their tangibility, ease of use, and inexpensive production. Cards are used for different purposes such as idea generation, inspiration, engagement, empathy, and to overcome problems that appear in a design process (Wölfel and Merritt 2013).

**Table 1** Participatory methods for design for values

| Participatory methods for design for values | |
|---|---|
| Card methods | Card methods in PD are used to inspire and explore and express emotions and values. They often come with specific instructions on how to use them to engage with fellow co-designers and to facilitate communication with the designers. For an overview of card methods, see Wölfel and Merritt (2013) |
| Mapping methods | Mapping methods are used to holistically map, express, and explore local knowledge and to enable those who will be affected by the design to express their values and to be actively involved in the design of an artifact |
| Future workshop | A method for planning and enacting possible futures, through the involvement of various participants (stakeholders) and with the use of various phases, towards a joint proposal for change (Jungk and Müllert 1981) |
| Participatory prototyping | Participatory prototyping is the process of generating, evaluating, and concretizing design ideas with active involvement of the future users (Lim et al. 2008) |

In PD, cards are often customizable: co-designers can add new cards or change cards. Cards can also be created during a discussion, expressing the themes or issues, which then can be sorted on priority or concept. Card sorting is a widely used design method in which the co-designer(s) organizes the cards into categories or selects particular cards to visualize processes, express priorities, or inspire creative processes (Fig. 2).

We designed and used a set of *inspiration cards* (Halskov and Dalsgård 2006) for our design work with teenagers with chronic health challenges preparing for the transition to the adult hospital. The cards addressed four categories connected with the transition process: *important people*, *things*, *feelings*, and *skills*. The cards consisted of an image with a one- or two-word concept description. The teen co-designers selected the cards they deemed important in each of the three stages in the transition process: preparing for transition, transitioning, and after transition. We also provided them with "empty" cards – cards without images or texts – providing them the opportunity to make their own inspiration cards and to add them to the stock of cards. In the process of selecting and organizing the cards and explaining their choices, the teens engaged with what was important to them. They visualized and verbalized what was important for them to know, to do, to feel, or to experience in the context of the transition process. The method facilitated the emergence of values in a more spontaneous and comprehensive manner and enabled better insight in the values at stake for themselves and for the designers. The inspiration cards are used to explore and engage with values early in the use-centered design cycle Fig. 1 (because it refers to the use-centered design cycle).

## Mapping Methods

Mapping methods are used to express and explore local knowledge. The resulting maps give holistic visualizations of geographies, contexts, life histories, workflows,


**Fig. 2** Inspiration cards (Photo: KULU)

or connections and relationships of the co-designers (Lanzara and Mathiassen 1985). Examples are context mapping (Visser et al. 2005) and cognitive mapping (Goodier and Soetanto 2013).

The *cartographic mapping method* is a method for making peoples' doings and activities in their everyday and working lives visible (Elovaara and Mörtberg 2010). The method is collective, simple, and cheap, and the material is familiar to all participants. The participants do not need to make any preparations in advance. The simplicity of the method encourages participants to start telling their stories and visualize their work or activities in everyday life and working life. The facilitators of the mapping activity prepare the workshop by collecting images of people and things and other materials such as yarn, pencils, colored paper, post-it notes, and large sheets to paste the images on. The workshop starts with a presentation of the method and the chosen topic. The room is arranged like a design space with a variety of materials. The participants are asked to take one sheet and to choose an image that presents them. The participants then start to include images of people, of things, and of their own activities (see Fig. 3). During the mapping activities, the facilitators ask the participants to clarify their choice of images and the connections between people and things pasted on the sheets. After the maps have been created,

**Fig. 3** Cartographic mapping

the participants are asked to share their stories with each other. The mapping activity initiates and contributes to a process in which values become expressed and materialized through the visualizations, the informal interviews, and the participants' oral stories of their cartographies. These stories can then be translated with the use of other methods and techniques, such as storyboards or mock-ups.

The method was developed and used in our research project *From government to e-government: skills, gender, technology, and learning* (Elovaara et al. 2006; Elovaara and Mörtberg 2010; Mörtberg et al. 2010). The purpose was to examine how the civil servants' skills and knowledge could be integrated in the design of the so-called e-service society. The method has also been used in other design projects, such as projects with an aim to design with and for elderly in their everyday lives.

### Future Workshop

The *future workshop* method was originally developed by Robert Jungk and Norbert Müllert (1981) to involve citizens in public decision-making processes, such as city planning. The method's basic principle, to enable participants to have a say, made it also of interest to PD (Kensing and Madsen 1991). The method's phases are preparation, critique, fantasy, realization, and follow-up. In the *preparation* phase the room is designed to create a welcoming atmosphere, and the method, the theme, and the facilitator are introduced. *Critique* of the current situation is generated in the next phase. Post-it notes can be used to write down

keywords, and these can then be pasted on a wall to make them visible for all. When a rich image of the current situation is generated, the keywords are organized into categories followed by a prioritization. In the *fantasy* phase, ideas are generated – a brainstorming without restrictions – and the critique is turned into something positive. Post-it notes can be used in a similar way as in the critique phase. Before the participants move to the *realization* phase, the generated ideas are analyzed and prioritized. In the realization phase, the participants review their visions and ideas and discuss the possibility for implementation. This is also a collaborative activity with the aim to create a joint proposal – an action plan, for how to change the current situation. A *follow-up* activity is also included in the method.

Future workshops were used in one of our projects with the aim to implement an automatic planning system in a handheld computer in a home care practice (Jansson 2007; Jansson and Mörtberg, 2011). A range of occupational groups involved in home care of elderly were invited to participate in the future workshops, in order to enable all stakeholders to be actively involved and to define their demands in the design and implementation of a digital artifact and new work practices. Through the involvement of various stakeholders, multiple values and views emerged out of the activities. This became visible during the generation of critique and visions, the systematization of the suggestions, their prioritization, and finally also when the participants consider possibilities for realization. Future workshops for technology design produce action plans as outcome, which form the first step in the materialization of requirements in the use-centered design cycle (see Fig. 1).

### Participatory Prototyping

Participatory prototyping, also called cooperative prototyping (Bødker and Grønbæk 1991) or collaborative prototyping, is one of the most important methods used in PD. Participatory prototyping is the activity in the design process in which values are translated into requirements and become materialized in a designed object. Prototypes can differ in material (paper, digital), resolution (mock-ups, low and high fidelity), and scope: they can be used to focus on a particular dimension of a design idea, which enables a more in-depth exploring (Lim et al. 2008).

In Participatory Design, prototyping creates a shared design space for designers and co-designers in which tensions between "what is" and "what could be" are explored through enacting scenarios, such as use situations (use-before-use). Prototyping can also provoke hindrances and new possibilities (Brodersen et al. 2008) or opportunities and dilemmas (Hillgren et al. 2011). Prototyping is an iterative process, often evolving from a low-fidelity prototype to a high-fidelity prototype with all the specifications of the finished product.

Several prototyping techniques can be used in one design process. In the KULU project, a design project with teenagers with chronic health challenges, we used sketches, paper mock-ups, as well as digital prototypes in the design of patient social media and mobile applications (KULU 2014). Our concern with the digital prototypes was that they might look too "finished" and thus present a solution before the design problem was fully explored. On the other hand, the digital

prototypes showed the teens a concrete rendering of their values and wishes, which encouraged them to engage with the prototypes, thus continuing the design process.

Privacy is one of the central values in the design projects with the teenagers. Research on teenagers with chronic health challenges and privacy in Canada has shown that most teens separate their identity as a patient from their (online) identity as a teenager and that this affects their online privacy behavior (van der Velden and El Emam 2013; van der Velden and Machniak 2014). During the prototyping activities with Norwegian teenage patients, a similar understanding of privacy emerged and materialized in different design projects. One example was the calendar function in a mobile application (app) for patient self-management. All teenage co-designers used the calendar function on their smartphone, but all wanted to have another calendar function within the new app. In a participatory prototyping session, it became clear that there was no consensus among the teenage co-designers on the use of the calendar function in the app. Some wanted to use it solely for medical reminders and appointments, while others wanted to use it for both medical and nonmedical reminders and appointments. Designers and co-designers agreed on a new prototype with colors and audio to differentiate between medical (orange) and nonmedical (green) appointments and reminders (Aasen 2014), and thus it provided the co-designers with different use scenarios for future use: medical-only, mixed medical/nonmedical, or both medical and nonmedical but not mixed (Fig. 4 is an example of a mixed medical/nonmedical calendar). The participatory prototyping session became a space for engaging with multiple understandings of privacy, which were concretized in a new calendar prototype, which was flexible enough to support the three use scenarios.

## Open Issues and Future Work

Genuine participation and power relations are two of the main issues in Participatory Design. There is no blueprint for addressing these challenges, but an ongoing critical and reflexive engagement with PD's guiding principles and design practices (e.g., Bratteteig and Wagner 2014) may facilitate ethical practices. It is here important to remember that not only people but also the structural arrangement in a design project, such as the allocation of resources or the choice of methods, affects the power to decide (ibid.). Reflexive and critical approaches become especially important in design projects with vulnerable groups, such as children, patients, and elderly. Several ethical heuristics for PD projects have been developed, which can be applied before and, iteratively, during a design process (e.g., Lahti et al. 2012; Phelan and Kinsella 2013; Read et al. 2013; Robertson and Wagner 2012, p. 82).

Another open issue is the political dimension in PD. Many industry-based projects lack an analysis of the larger socioeconomic and cultural structures in which we design, produce, and use artifacts, as the focus is on the production of consumer goods (Bergvall-Kåreborn and Ståhlbrost 2008). The result is the risk of a more instrumental use of the co-designers in a design project, as Beck already observed in 2002: "A politicised agenda for PD would need to centrally address, then, the

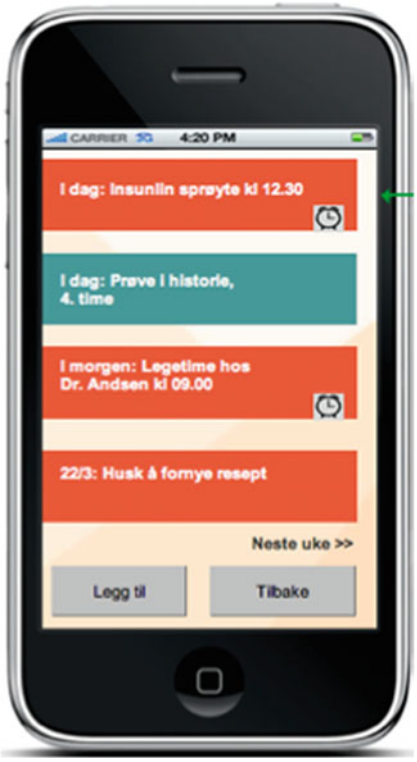



**Fig. 4** Calendar function (Aasen 2014)

legitimacy of anyone not only to propose solutions, but to suggest what the problems are" (Beck 2002, p. 83). Fry (2009, 2011), concerned with the issue of sustainment, argues that PD could strengthen the voice of the *common good* and delegitimize *consumer democracy*. One way to address these concerns is to let the future generations *have a say* in today's PD projects (van der Velden 2014). Bringing a multigenerational perspective into PD may result in frontloading certain values, such as sustainment, and the design of new methods or redesign of existing methods, such as the *future workshop*. Also the insights of other design approaches, such as Design Futuring (Fry 2009) and Metadesign (Fischer et al. 2004; Wood 2007), can strengthen PD's political dimension.

Future work in Participatory Design may further explore the relation between values and PD methods. Methods play a central role in creating a space for the emergence of values and in engaging designers and co-designers in the expression and exploration of these values. Overviews of participatory methods are helpful (e.g., Sanders et al. 2010; Wölfel and Merritt 2013) but focus mainly on functionality and application area of methods. Reviews of participatory methods based on how they may support the exploration, engagement with, and materialization of values may provide participatory designers with an important tool for strengthening their value practices.

## Concluding Remarks

Participatory Design's principal values, *democracy* and *participation*, make it an important methodology for Design for Values. These values shape the design process, resulting in: a) a genuine engagement with the people who will be using the outcome of the process and b) the use of design methods that focus on creating a shared design space in which designer and co-designer values are expressed and become materialized in a product or service.

We have argued that the design process in a PD project is as important as its outcome. During the design process, other goals are accomplished, such as mutual learning, reflection, and skill acquisition, which have a value that is independent of the final outcome of the process. Participatory Design's guiding principles prevent the design process to become a purely instrumental phase for eliciting user input for a final product. By facilitating genuine participation, engaging the experiences, skills, needs, and values of the co-designers, the design process becomes a space in which alternative visions about technology are envisioned and anticipated. Through the creative use of participatory methods, the Participatory Design process enables use-before-use scenarios, in which co-designers can try out or engage with the result(s) before it is produced or implemented. The result of this process is a product or service that matters, because the co-designers' values and needs have become materialized in the design.

## Cross-References


- ▶ Design Methods in Design for Values
- ▶ Design for the Value of Presence
- ▶ Design for the Values of Democracy and Justice
- ▶ Value Sensitive Design: Applications, Adaptations, and Critiques


## References


Aasen N (2014) Transisjonsapp - Ansvar for egen helse. Master thesis, University of Oslo, Oslo

Akrich M (1992) The description of technical objects. In: Bijker W, Law J (eds) Shaping technology. MIT Press, Cambridge, pp 205–224

Albrechtslund A (2007) Ethics and technology design. Ethics Inf Technol 9(1):63–72

Bath C (2009) Searching for methodology: feminist technology design in computer science. Online Proc. gender & ICT. Retrieved from http://www.informatik.uni-bremen.de/soteg/gict2009/proceedings/GICT2009_Bath-geloescht.pdf

Beck E (2002) P for political: participation is not enough. Scand J Inf Syst 14(1)

Bergvall-Kåreborn B, Ståhlbrost A (2008) Participatory design: one step back or two steps forward? In: Proceedings of the tenth anniversary conference on participatory design 2008. Indiana University, Indianapolis, pp 102–111

Bjerknes G, Bratteteig T (1995) User participation and democracy: a discussion of Scandinavian research on system development. Scand J Inf Syst 7:73–73

Blomberg J, Karasti H (2012) Ethnography: positioning ethnography within participatory design. In: Routledge international handbook of participatory design. Routledge, New York, pp 86–116
Blomberg JL, Giacomi J, Mosher A, Swenton-Hall P (1993) Ethnographic field methods and their relation to design. In: Participatory design: principles and practices. Lawrence Erlbaum Associates, Hillsdale, pp 123–156
Bødker S (1996) Creating conditions for participation: conflicts and resources in systems development. Hum-Comput Interact 11(3):215–236
Bødker S, Grønbæk K (1991) Cooperative prototyping: users and designers in mutual activity. Int J Man–Machine Stud 34(3):453–478
Bødker K, Kensing F, Simonsen J (2004) Participatory IT design: designing for business and workplace realities. MIT Press, Cambridge
Braa J, Sahay S (2012) Health information systems programme: participatory design within the HISP network. In: Routledge international handbook of participatory design. Routledge, New York, pp 235–256
Brandt E, Binder T, Sanders EB-N (2012) Tools and techniques: ways to engage telling, making, and enacting. In: Routledge international handbook of participatory design. Routledge, New York, pp 145–181
Bratteteig T (2004) Making change. Dealing with relations between design and use. Dr. Philos dissertation, Department of Informatics, University of Oslo
Bratteteig T, Stolterman E (1997) Design in groups – and all that jazz. Comput Des Context 289–316
Bratteteig T, Wagner I (2014) Disentangling participation: power and decision-making in participatory design. Springer, Heidelberg
Bratteteig T, Bødker K, Dittrich Y, Mogensen PH, Simonsen J (2012) Methods: organising principles and general guidelines for participatory design projects. In: Routledge international handbook of participatory design. Routledge, New York, pp 117–144
Brey P (2010) Values in technology and disclosive computer ethics. Camb Handb Inf Comput Ethics 41–58
Brodersen C, Dindler C, Iversen OS (2008) Staging imaginative places for participatory prototyping. CoDesign 4(1):19–30
Christiansen E (2014) From "ethics of the eye" to "ethics of the hand" by collaborative prototyping. J Inf Commun Ethics Soc 12(1):3–9
Dindler C, Iversen OS (2007) Fictional inquiry – design collaboration in a shared narrative space. CoDesign 3(4):213–234
Ehn P (1988) Work-oriented design of computer artifacts. Arbetslivscentrum, Stockholm
Ehn P (2008) Participation in design things. In: Proceedings of the tenth anniversary conference on participatory design 2008. Indiana University, Indianapolis, pp 92–101
Ehn P, Badham R (2002) Participatory design and the collective designer. PDC 1–10
Ehn P, Sandberg Å (1979) Systems development: on strategy and ideology. Data (4):1–8
Elovaara P, Mörtberg C (2010) Cartographic mappings: participative methods. In: Proceedings of the 11th biennial participatory design conference. ACM, New York, pp 171–174
Elovaara P, Igira FT, Mörtberg C (2006) Whose participation? whose knowledge?: exploring PD in Tanzania-Zanzibar and Sweden. In: Proceedings of the ninth conference on participatory design: expanding boundaries in design, vol 1. ACM, New York, pp 105–114
Finken S, Stuedahl D (2008) Silence' as an analytical category for PD. In: Proceedings of the tenth anniversary conference on participatory design 2008. Indiana University, Indianapolis, pp 170–173
Fischer G, Giaccardi E, Ye Y, Sutcliffe AG, Mehandjiev N (2004) Meta-design: a manifesto for end-user development. Commun ACM 47(9):33–37. doi:10.1145/1015864.1015884
Flyvbjerg B (2001) Making social science matter: why social inquiry fails and How it Can succeed again. Cambridge University Press, Cambridge, UK
Friedman B, Kahn PH (2003) In: Jacko JA, Sears A (eds) The human-computer interaction handbook. Erlbaum Associates, Hillsdale

Fry T (2009) Design futuring: sustainability, ethics and new practice. Berg/Macmillan, Oxford
Fry T (2011) Design as politics. Berg, New York
Goodier CI, Soetanto R (2013) Building future scenarios using cognitive mapping. J Maps 9(2):203–217
Greenbaum JM, Kyng M (1991) Design at work: cooperative design of computer systems. Lawrence Erlbaum Associates, Hillsdale
Gregory J (2003) Scandinavian approaches to participatory design. Int J Eng Educ 19(1):62–74
Halloran J, Hornecker E, Stringer M, Harris E, Fitzpatrick G (2009) The value of values: resourcing co-design of ubiquitous computing. CoDesign 5(4):245–273
Halskov K, Dalsgård P (2006) Inspiration card workshops. In: Proceedings of the 6th conference on designing interactive systems. ACM, New York, pp 2–11
Haraway D (1988) Situated knowledges: the science question in feminism and the privilege of partial perspective. Fem Stud 14(3):575–599
Haraway D (2003) The companion species manifesto: dogs, people, and significant otherness. Prickly Paradigm Press, Chicago
Hillgren P-A, Seravalli A, Emilson A (2011) Prototyping and infrastructuring in design for social innovation. CoDesign 7(3–4):169–183
Hussain S, Sanders EB-N, Steinert M (2012) Participatory design with marginalized people in developing countries: challenges and opportunities experienced in a field study in Cambodia. Int J Des 6(2)
Ihde D (1999) Technology and prognostic predicaments. AI Soc 13(1–2):44–51
Iversen OS, Halskov K, Leong TW (2012) Values-led participatory design. CoDesign 8(2–3):87–103
Jansson M (2007) Participation, knowledges, and experiences: design of IT- systems in e-home health care. Doctoral thesis, Luleå Technical University. Retrieved from http://epubl.ltu.se/1402-1544/2007/56/index-en.html
Jansson M, Mrtberg C (2011) A Cup of Coffee: Users' needs and experiences of ICT in homecare. In: Human-Centered Design of E-Health technologies: Concepts, Methods, and Applications. IGI Global, Hershey, pp 253–271
Jungk R, Müllert NR (1981) Zukunftswerkstätten, wege zur wiederbelebung der demokratie (future workshops: ways to revive democracy). Goldman, Munchen
Kanstrup AM, Christiansen E (2006) Selecting and evoking innovators: combining democracy and creativity. In: Proceedings of the 4th Nordic conference on human-computer interaction: changing roles. ACM, New York, pp 321–330
Karasti H (2001) Increasing sensitivity towards everyday work practice in system design. University of Oulu, Oulu
Karasti H (2003) Gendered expertise and participation in systems design. In: How to make a difference?: information technology, transnational democracy and gender. Luleå University of Technology, Luleå, pp 29–49
Keinonen T (2010) Protect and appreciate – notes on the justification of user-centered design. Int J Des 4(1):17–27
Kensing F, Greenbaum J (2012) Heritage: having a say. In: Routledge international handbook of participatory design. Routledge, New York, pp 21–36
Kensing F, Madsen KH (1991) Generating visions: future workshops and metaphorical design. In: Design at work: cooperative design of computer systems. Lawrence Erlbaum Associates, Hillsdale, pp 155–168
KULU (2014) Kul teknologi for unge med langvarige helseutfordringer. http://www.kulu.no/english.php. Retrieved 30 May 2014
Lahti A, Naraha S, Svensson J, Wärnestål P (2012) Ethical heuristics – a tool for applying ethics in user-involved IS projects. Presented at the nordic contributions in IS research: third scandinavian conference on information systems, SCIS 2012 Sigtuna Sweden. Proceedings retrieved from http://hh.diva-portal.org/smash/record.jsf?pid=diva2:548656., 17–20 Aug 2012

Lanzara GF, Mathiassen L (1985) Mapping situations within a system development project. Inf Manag 8(1):3–20
Latour B (1991) Technology is society made durable. In: Law J (ed) A sociology of monsters: essays on power, technology, and monsters. Routledge, London, pp 103–131
Letondal C, Mackay WE (2004) Participatory programming and the scope of mutual responsibility: balancing scientific, design and software commitment. In: Proceedings of the eighth conference on participatory design: artful integration: interweaving media, materials and practices, vol 1. ACM, New York, pp 31–41
Lim Y-K, Stolterman E, Tenenberg J (2008) The anatomy of prototypes: prototypes as filters, prototypes as manifestations of design ideas. ACM Trans Comput-Hum Interact 15(2):7:1–7:27
Luck R (2003) Dialogue in participatory design. Des Stud 24(6):523–535
Luck R (2007) Learning to talk to users in participatory design situations. Des Stud 28(3):217–242
Manders-Huits N, van den Hoven J (2009) The need for a value-sensitive design of communication infrastructures. In: Sollie P, Düwell M (eds) Evaluating new technologies. Springer, Dordrecht, pp 51–60
Manders-Huits N, Zimmer M (2009) Values and pragmatic action: the challenges of introducing ethical intelligence in technical design communities. Int Rev Inf Ethics 10(2):37–45
Markussen R (1996) Politics of intervention in design: feminist reflections on the Scandinavian tradition. Ai Soc 10(2):127–141
Morgall J, Vedel G (1985) Office automation: the case of gender and power. Econ Ind Democr 6(1):93–112
Mörtberg C, Stuedahl D (2005) Silences and sensibilities: increasing participation in IT design. In: Proceedings of the 4th decennial conference on critical computing: between sense and sensibility. ACM, New York, pp 141–144
Mörtberg C, Bratteteig T, Wagner I, Stuedahl D, Morrison A (2010) Methods that matter in digital design research. In: Wagner I, Bratteteig T, Stuedahl D (eds) Exploring digital design. Springer, London, pp 105–144
Mouffe C (1993) The return of the political. Verso, London
Nygaard K (1974) Planlegging, styring og databehandling: Grunnbok for fagbevegelsen Del 2 Datamaskiner, systemer og språk. Tiden Norsk Forlag, Oslo
Nygaard K, Bergo OT (1975) The trade unions – new users of research. Pers Rev 4(2):5–10
Phelan SK, Kinsella EA (2013) Picture this ... safety, dignity, and voice – ethical research with children practical considerations for the reflexive researcher. Qual Inq 19(2):81–90
Pratt ML (1998) Arts of the contact zone. Negot Acad Lit: Teach Learn Across Lang C 91:171
Puri SK, Byrne E, Nhampossa JL, Quraishi ZB (2004) Contextuality of participation in IS design: a developing country perspective. In: Proceedings of the eighth conference on participatory design: artful integration: interweaving media, materials and practices, vol 1. ACM, New York, pp 42–52
Read JC, Horton M, Sim G, Gregory P, Fitton D, Cassidy B (2013) CHECk: a tool to inform and encourage ethical practice in participatory design with children. In: CHI'13 extended abstracts on human factors in computing systems. ACM, New York, pp 187–192
Redström J (2006) Towards user design? On the shift from object to user as the subject of design. Des Stud 27(2):123–139
Redström J (2008) RE:definitions of use. Des Stud 29(4):410–423
Robertson T, Wagner I (2012) Ethics: engagement, representation and politics-in-action. In: Routledge international handbook of participatory design. Routledge, New York, pp 64–85
Sanders EBN, Brandt, E, Binder T (2010) A framework for organizing the tools and techniques of participatory design. In: Proceedings of the 11th biennial participatory design conference (pp. 195–198). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1900476
Shapiro D (2010) A modernised participatory design? Scand J Inf Syst 22(1):69–76
Simonsen J, Robertson T (2012) Routledge international handbook of participatory design. Routledge, New York

Spinuzzi C (2005) The methodology of participatory design. Techl Commun 52(2):163–174
Star SL, Strauss A (1999) Layers of silence, arenas of voice: the ecology of visible and invisible work. Comput Support Coop Work (CSCW) 8(1–2):9–30
Steen M (2013) Virtues in participatory design: cooperation, curiosity, creativity, empowerment and reflexivity. Sci Eng Ethics 19(3):945–962
Stuedahl D (2004) Forhandlinger og overtalelser: Kunnskapsbygging pÍ tvers av kunnskapstradisjoner i brukermedvirkende design av ny IKT. University of Oslo, Oslo
Suchman L (1995) Making work visible. Commun ACM 38(9):56–64
Suchman L (2002) Located accountabilities in technology production. Scand J Inf Syst 14(2):91–106
Suchman L (2007) Human-machine reconfigurations: plans and situated actions, 2nd edn. Cambridge University Press, Cambridge
Telier A, Binder T, De Michelis G, Ehn P, Jaccuci G, Linde P, Wagner I (2011) Design things. MIT Press, Cambridge, MA
Van den Hoven J (2007) ICT and value sensitive design. In: Goujon P, Lavelle S, Duquenoy P, Kimppa K, Laurent V (eds) The information society: innovation, legitimacy, ethics and democracy in honor of Professor Jacques Berleur S.J. Springer, New York, pp 67–72. Retrieved from http://link.springer.com/chapter/10.1007/978-0-387-72381-5_8
Van der Velden M (2010) Design for the contact zone. Knowledge management software and the structures of indigenous knowledges. In: Sudweeks F, Hrachovec H, Ess C, Sudweeks F, Hrachovec H, Ess C (eds) Cultural attitudes towards technology and communication 2010 proceedings of the seventh international conference on cultural attitudes towards technology and communication Vancouver, Canada, 15–18 June 2010. School of Information Technology, Murdoch University, Murdoch, pp 1–18. Retrieved from http://sammelpunkt.philo.at:8080/2002/
Van der Velden M (2014) Re-politicising participatory design: what we can learn from fairphone. In: Ninth international conference on culture, technology, and communication. Oslo. Retrieved from http://philo.at/ocs2/index.php/oslo14/ctnewd14/paper/view/295
Van der Velden M, El Emam K (2013) "Not all my friends need to know": a qualitative study of teenage patients, privacy, and social media. J Am Med Inform Assoc: JAMIA 20(1):16–24
Van der Velden M, & Machniak M (2014) Colourful privacy: designing visible privacy settings with teenage hospital patients. Presented at the ACHI 2014, The seventh international conference on advances in computer-human interactions, pp 60–65. Retrieved from http://www.thinkmind.org/index.php?view=article&articleid=achi_2014_3_30_20220
Vehviläinen M (1997) Gender, expertise and information technology. University of Tampere, Department of Computer Science, Tampere
Verbeek P-P (2006) Materializing morality design ethics and technological mediation. Sci Technol Hum Value 31(3):361–380
Verbeek P-P (2011) Moralizing technology: understanding and designing the morality of things. University of Chicago Press, Chicago
Visser FS, Stappers PJ, van der Lugt R, Sanders EB-N (2005) Contextmapping: experiences from practice. CoDesign 1(2):119–149
Wagner I (1993) A Web of fuzzy problems: confronting the ethical issues. Commun ACM 36(6):94–101
Winschiers-Theophilus H, Bidwell NJ, Blake E (2012) Altering participation through interactions and reflections in design. CoDesign 8(2–3):163–182
Wölfel C, Merritt T (2013) Method card design dimensions: a survey of card-based design tools. In: Kotzé P, Marsden G, Lindgaard G, Wesson J, Winckler M (eds) Human-computer interaction – INTERACT 2013. Springer, Berlin/Heidelberg, pp 479–486
Wood J (2007) Designing for micro-utopias; thinking beyond the possible. Ashgate. Retrieved from http://research.gold.ac.uk/326/

# Technology Assessment and Design for Values

Armin Grunwald

## Contents



**Abstract**

Technology assessment (TA) constitutes a scientific and societal response to problems at the interface between technology and society. It is a field that has arisen against the background of various experiences concerning the unintended and often undesirable side effects of science, technology, and societal technicization. This chapter provides an overview of the history, motivations, objectives, and present status of TA. Elements of the governance of technology are discussed in order to identify appropriate constellations where knowledge and orientation provided by TA could be used to improve decision making. There are three major branches of TA: TA as policy advice (e.g., to parliaments), TA in public debate (e.g., by participatory measures), and TA for shaping

A. Grunwald (✉)
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: armin.grunwald@kit.edu



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_4

technology directly (e.g., by constructive technology assessment or by *Leitbild* assessment). In all of these branches, TA is considering relations between technology and values. In particular, insofar as TA is involved in processes of shaping technology directly, there is a close neighborhood with Design for Values.

**Keywords**
Ambivalence of technology • Side effects • Innovation • Risk • Technology conflicts • Policy advice

## Introduction and Overview

Technology assessment (TA) constitutes an interdisciplinary research field aiming at, generally speaking, providing knowledge for better-informed and well-reflected decisions concerning new technologies. Its initial and still valid motivation is to provide answers to the emergence of unintended and often undesirable side effects of science, technology, and technicization (Bechmann et al. 2007). TA shall add reflexivity to technology governance by integrating any available knowledge on the side effects at an early stage in decision-making processes, by supporting the evaluation of technologies and their impact according to societal values and ethical principles, by elaborating strategies to deal with the uncertainties that inevitably arise, and by contributing to constructive solutions of societal conflicts. Values play a crucial role in all of these fields. There are three branches of TA addressing different targets in the overall technology governance:

1. TA has initially been conceptualized as *policy advice* (Bimber 1996), and still many TA activities are located in this field (Grunwald 2009a). The objective is to support policymakers in addressing the abovementioned challenges by implementing political measures such as adequate regulation (e.g., the precautionary principle), sensible research funding, and strategies toward sustainable development involving appropriate technologies. In this mode of operation, TA does not *directly* address technology development but considers the *boundary conditions* of technology development and use.
2. It became clear during the past decades that citizens, consumers and users, actors of civil society, stakeholders, the media, and the public are also engaged in technology governance in different roles. Participatory TA developed approaches to involve these groups in different roles at different stages in technology governance (Joss and Belucci 2002).
3. A third branch of TA is related more directly to concrete technology development and engineering. Departing from analyses of the genesis of technology made in the framework of social constructivism (Bijker et al. 1987) the idea of *shaping technology* due to social expectations and values came up and motivated the development of several approaches such as constructive TA (CTA) or social shaping of technology (Yoshinaka et al. 2003). They all aim at increasing

reflexivity in technology development and engineering by addressing the level of concrete products, systems, and services, going for a "better technology in a better society" (Rip et al. 1995).

All these branches of TA have to deal with social values and ethical reflection. In this chapter, I will introduce briefly all these three branches but will then focus on the third one in order to identify sources of the idea of Design for Values. At the beginning the overall motivations of TA will be presented (section "Motivations of Technology Assessment") and a brief insight into technology governance will be given (section "Elements of Technology Governance"). Section "Technology Assessment: Adding Reflexivity to Technology Governance" is dedicated to explaining the very idea of TA for increasing reflexivity in technology governance which will be done briefly with respect to the three branches of TA mentioned above. More in depth I will then look at approaches of TA which explicitly address technology development and design, such as the constructive TA and the social shaping of technology approach (see section "TA Approaches Relevant for Design for Values").

## Motivations of Technology Assessment[1]

In the twentieth century, the importance of science and technology in almost all areas of society (touching on economic growth, health, the army, etc.) has grown dramatically. Concomitant with this increased significance, the consequences of science and technology for society and the environment have become increasingly serious. Technological progress alters social traditions, fixed cultural habits, relations of humans and nature, collective and individual identities, and concepts of the self while calling into question traditional ethical norms. Decisions concerning the pursual or abandonment of various technological paths, regulations and innovation programs, new development plans, or the phasing out of lines of technology often have far-reaching consequences for further development. They can influence competition in relation to economies or careers, trigger or change the direction of flows of raw materials and waste, influence power supplies and long-term security, create acceptance problems, fuel technological conflict, challenge value systems, and even affect human nature (Habermas 2001).

Since the 1960s also adverse effects of scientific and technical innovations became obvious; some of them were of dramatic proportions: accidents in technical facilities (Chernobyl, Bhopal, Fukushima), threats to the natural environment (air and water pollution, ozone holes, climate change), negative health effects as in the asbestos case, social and cultural side effects (e.g., labor market problems caused by productivity gains), and the intentional abuse of technology (the attacks on the

[1]This section follows the introduction to TA given in Grunwald (2009a). Some paragraphs were taken from that chapter and adapted in a shortened way.

World Trade Center). The experience with such unexpected and serious impacts of technology is central to TA's motivation. Indeed, in many cases, it would have been desirable to have been warned about the disasters in advance, either to prevent them or to be in a position to undertake compensatory measures. This explains why the methodologically quite problematic term "early warning" with regard to technological impacts (Bechmann 1994) has always had a prominent place in TA discussions from the very beginning (Paschen and Petermann 1992, p. 26).

Early warning is a necessary precondition to make societal and political *precautionary action* possible: how can a society which places its hopes and trust in innovation and progress, and must continue to do so in the future, protect itself from undesirable, possibly disastrous side effects, and how can it preventatively act to cope with possible future adverse effects? Classic problems of this type are, for example, the use and release of new chemicals – the catastrophic history of asbestos use being a good example (Gee and Greenberg 2002) – and dealing with artificial or technically modified organisms (for further examples, cf. Harremoes et al. 2002). In order to be able to cope rationally with these situations of little or no certain knowledge of the effects of the use of technology, prospective precautionary research and corresponding procedures for societal risk management are required, for instance, by implementing the precautionary principle (von Schomberg 2005; Grunwald 2008).

Parallel to these developments, broad segments of Western society were deeply unsettled by the "Limits of Growth" (Club of Rome) in the 1970s which, for the first time, addressed the grave environmental problems perceived as a side effect of technology and technicization. The optimistic pro-progress assumption that whatever was scientifically and technically new would definitely benefit the individual and society was questioned. As of the 1960s deepened insight into technological ambivalence led to a crisis of orientation in the way society dealt with science and technology. Without this (persistent!) crisis TA would presumably never have developed.

New and additional motivations entered the field of TA over the past decades, leading more and more to a shift toward "shaping technology" according to social values (and, therefore, building a bridge to the idea of Design for Values):

- Issues of democracy and technocracy or of democratizing technology (von Schomberg 1999): from the 1960s on, there are concerns that the scientific and technological advance could threaten the functioning of democracy because only few experts were capable of really understanding the complex technologies. The technocracy hypothesis was born painting a picture of a future society where experts would make the decisions on the basis of their own value systems. One of the many origins of TA is to counteract this possibility and to enable and empower society to take active roles in democratic deliberation (Joss and Belucci 2002; Grunwald 2003).
- The experience of technology conflicts, of legitimacy deficits, and of little acceptance of some decisions on technology motivated TA to think about more socially compatible technology. The very idea was to design technology

according to social values – and if this would succeed, so the hope was, problems of rejection or nonacceptance would no longer occur. This line of thought seems to be one of the sources of Design for Values (see section "TA Approaches Relevant for Design for Values").

- In the past decade the innovation problems of Western societies influenced also the motivations and driving forces of TA. TA was considered part of regional and national innovation systems (Smits and den Hertog 2007) which could contribute to "responsible innovation" and "responsible development" (Siune et al. 2009) by taking into account not only technical and economical but also social and ethical aspects.
- Shift in the societal communication on new and emerging science and technology (NEST): techno-visionary sciences such as nanotechnology, converging technologies, and synthetic biology entered the arena. Visions and metaphors mark the expected revolutionary advance of science in general and became an important factor in societal debates. To provide for more rationality, reflexivity, and transparency in these debates, vision assessment was proposed (Grunwald 2009b) as a new TA tool addressing not directly the assessment of *technologies* but rather the assessment of *visions* (Grin and Grunwald 2000). In particular, vision assessment aims at reconstructing normative elements of the visions under consideration including inherent values.
- Finally, recent debates around ethics in the field of biomedicine (e.g., stem cell research, xenotransplantation, and reproduction medicine) led to a convergence of applied ethics and TA in some regard and complemented the agenda of TA by issues of bioethics and medical ethics.

Compared to the initial phase of TA, a considerable increase of its diversity and complexity can be observed. In modern TA, it is often a question not only of the consequences of individual technologies, products, or plants but also frequently of complex conflict situations between enabling technologies, innovation potentials, fears and concerns, patterns of production and consumption, lifestyle and culture, and political and strategic decisions. The challenge of "responsible innovation" (Siune et al. 2009) can be seen as a core to which all of these research and assessment branches contribute, setting out from different premises, using different perspectives, and applying different TA methodologies (see section "TA Approaches Relevant for Design for Values").

## Elements of Technology Governance

Technology is being shaped and influenced in a complex process of technology governance (Aichholzer et al. 2010). TA shall "make a difference" in this process – and in order to be "really" able to make a difference, TA must have sound knowledge about the processes of technology development and diffusion, about the pathways from research to innovation, about social integration and adaptation of new technology, about influencing and decisive factors in these processes, and so

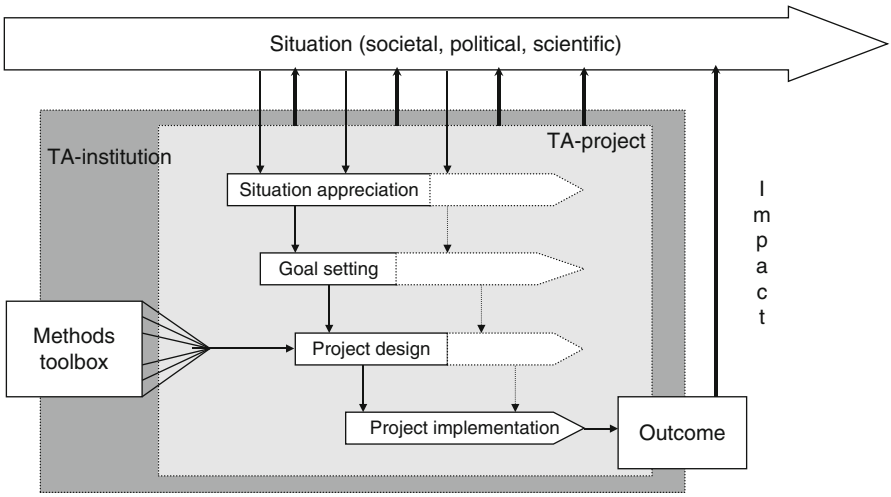



**Fig. 1** TA influencing the ongoing societal situation by concrete TA projects continuously keeping track with developments at the societal level (Decker and Ladikas 2004)

forth. In the TAMI project (Decker and Ladikas 2004), the very idea of TA in making a difference was conceptualized in the following way (Fig. 1):

A complete picture of all the possible entry points of technology assessment would require a complete theory of technology in society. Such a theory would have to include theories of the origin and genesis of technology, the route technology takes through society during the phase of its utilization, and the manner in which society deals with a technology after its use is discontinued – this would probably be no less than a comprehensive theory of society which is not available. Regarding this situation, I will restrict myself to briefly describing important elements of the overall technology governance and of relevant actors.

Governance of science (Siune et al. 2009) as well as the governance of technology has become much more diverse and complex over the past decades. While in earlier times (in the "classical mode" of TA, cp. Grunwald 2009a) a strong role of the state was supposed, nowadays much more actors and stakeholders are regarded as being influential on the development and use of new technologies: companies, consumers, engineers, nongovernmental organizations (NGO), stakeholders of different kinds, and citizens. Depending on their roles and occasions to take influence, the advice provided by TA could or should look different – in this sense the shift from "steering technology" to a "governance of technology" has had a major influence on TA. Theories of technology development and governance could provide orientation for TA whom to address and what to deliver.

However, the political system remains a major player since public research and technology policy create legitimate and binding decisions with (partially) high influence on technology. Policy consultation by TA can, for example, take place in the preparatory phase of legislation relevant to technology or even in the very early phases of opinion forming in the political parties. In the run-up to policy

decisions, it is possible for TA to carry out enlightenment by reflecting on possible consequences and impacts of technology on society and on the values touched (Grunwald 2003). This positioning of TA research and consultation affects all constellations in which state action influences technology including:

- Direct state-run or at least state-dominated technology development, for example, in the fields of space travel, military technology, and transportation infrastructure
- Indirect political influence on technology by means of programs promoting research and technology, for example, in materials science, on regenerative sources of energy, or in stem cell research
- Indirect political control of technology by setting boundary conditions such as environmental and safety standards, laws on privacy, or laws stipulating recycling
- The role of the state as a *user* of technology, e.g., with regard to the observance of sustainability standards (public procurement) and to its capability to create or support lead markets for innovative developments

TA gives advice to policymakers in all of these fields and to the involved organizations such as parliaments, governments, and authorities. An example is the *Office of Technology Assessment* at the German *Bundestag* (TAB: http://www.tab-beim-bundestag.de). TAB improves the legislature's information basis, in particular, of research- and technology-related processes of parliamentary discussion. TAB performs this mission in scientific independence (Grunwald 2006). Among its responsibilities are, above all, drawing up and carrying out TA projects, and – in order to prepare and to supplement them – observing and analyzing important scientific and technical trends, as well as societal developments associated with them (monitoring). The TAB is strictly oriented on the German *Bundestag's* and its committees' information and advice requirements. The subjects of the TAB's studies stem from all fields of technology and its applications such as energy, bio-, and genetic engineering; defense technologies; nanotechnology and materials research; space flight; medical technologies; and information and communication technologies.

The concrete development of technology and innovation, however, takes place primarily in the economy at market conditions. The shaping of technology by and in enterprises is operationalized by means of requirement specifications, project plans, and strategic entrepreneurial decisions. These in turn take place on the prescriptive basis of an enterprise's headline goals, general principles, plan goals, and self-understanding but also including assumptions about later consumers and users of the technology as well as future market conditions. Engineers and engineering scientists have influence on decisions at this level and are confronted in a special way with attributions of responsibility because of their close links with the processes of the development, production, utilization, and disposal of technology (Durbin and Lenk 1987; van Gorp 2005). Technology assessment became aware of the importance of this part of technology governance in the 1980s in the course of

the social constructivist movement leading to the slogan of "shaping technology" (Bijker and Law 1994).

The individual preferences of users and consumers of technical systems and products help determine the success of technology developments in two ways: first, by means of their purchasing and consumer behavior and, second (and less noted), by means of their comments in market research. The influence on technological development resulting from consumer behavior arises from the concurrence of the actual purchasing behavior of many individual persons. A well-known problem is, for example, that awareness of a problem with regard to the deficient environmental compatibility of certain forms of behavior – though definitely present – may not lead to a change in behavior. Technology assessment aims, in this field, at public enlightenment and information about consequences of consumer's behavior and at enabling and empowering individuals to behave more reflexively.

The course of technical development is also decided by public debates, above all by those in the mass media. Public discussion in Germany influenced, for example, political opinion on atomic energy, thus providing much of the basis for the decision in 2002 to phase out atomic energy quickly in that country and to return to this position after a short, more positive appraisal of nuclear energy within hours after the Fukushima disaster. Similarly, the public discussion about genetically modified organisms has influenced the regulatory attitude of the European Union and the official acceptance of the precautionary principle (von Schomberg 2005). This can also be recognized by the fact that different regulations were established in those countries in which the public debates were very different, such as in the USA. Many of the public debates conducted in the media have also influenced the policy framework with indirect influence on technology. Technology assessment agencies have become an actor also in this field by involving themselves in participatory processes that play an increasing role in political decision-making processes in many countries.

These brief remarks should give some insights into the complex nature of technology governance – implying a similar complexity of technology assessment aiming at giving advice to actors in the various fields of technology governance. In the following, three fields of importance in and for TA will be described in some more detail: (a) TA as policy advice, (b) TA supporting public debate, and (c) TA aiming at shaping technology directly.

## Technology Assessment: Adding Reflexivity to Technology Governance

Different TA approaches have been proposed and practiced responding to the societal context and to elements of technology governance, e.g., participative TA (Joss and Belucci 2002), constructive TA (CTA, Rip et al. 1995), interactive TA, TA relying on innovation systems research (Smits and den Hertog 2007), and others. On the one hand the differentiation is due to different questions each of them is suited to address; on the other it is due to different basic distinctions and

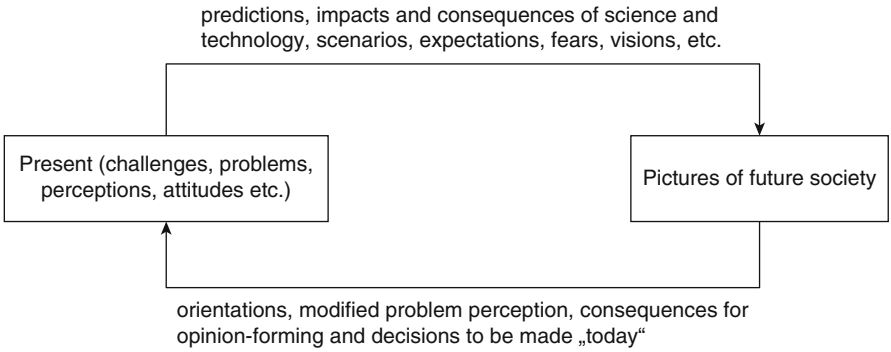



**Fig. 2** The decision-making cycle via future thinking (Grunwald 2011)

assumptions about technology governance which relate directly to images and models of the technological evolution, the role of the state or the market in modern societies, how shaping of technology should work in democracies, etc. In this section I would like to demonstrate that there is a common element of TA beyond its diversity: the impetus to increase reflexivity in technology governance.

The theory of "reflexive modernization" (Beck 1986) stated, among other works on theory of modern society, that information gained from reflection on the future is a commonly used means for facilitating the decision-making process. Prospective knowledge of consequences, prognoses of technical progress, expectations, and fears, as well as aims, serves to provide orientation *today* for pending decisions (cf. Fig. 2). Proceeding from *present-day* problem perceptions, grand challenges, and expectations, orientation *for today* is sought via the roundabout route of debates about the *future*.

To provide orientation by reflecting on futures is a highly ambitious undertaking. For this to succeed, the loop of decision making (see Fig. 2) may not be a vicious circle, i.e., an idle state of knowledge, but must demonstrate added value that provides orientation compared to the situation *prior to* entering the cycle (cp. Grunwald (2011) for the case of energy futures). TA is in charge of contributing to constructively meeting this challenge in different fields.

## TA for Policy Advice: The Case of Parliaments

Parliamentary TA is part of TA with a tradition of decades and with diverse forms of institutionalization (Vig and Paschen 1999; Cruz-Sastro and Sanz-Menendez 2004). It is about advising parliamentary actors within the frameworks based on the respective structures of the nation-state. In order to be able to analyze the role of parliamentary TA for technology governance, we have to take a closer look at the general role of the state in technology governance first. Obviously the provision of policy advice can be purposive, sensible, and effective only under the assumption

that political actors and institutions play at least a rather important role in the overarching processes of technology governance.

However, this precondition is controversial. Traditional nation-states frequently are regarded as having lost their scopes for steering other actors like industry, supranational institutions, actors of civil society, or informal and nonhierarchical processes. Their remaining role would only be the role as a moderator of societal processes of communication. There is some evidence in this position but it does not justify the very far-reaching conclusions. Also in modern and less hierarchically structured societies, the democratic state with its institutions and procedures remains the sole place to produce legitimized, generally binding decisions. Of course this also applies for decisions, which concern technology and which are binding for all (cf. Grunwald 2000a).

The undoubted fact that technology and innovation development is definitely mainly taking place in the industry under market conditions does not exclude or diminish the relevance of political influence on technology. In a thought experiment we could distinguish between different aspects of technological products or systems: aspects bound to political reasoning (environmental norms, safety regulations, technical standardizations, general statutory provisions, etc.) and aspects which could be delegated to market developments. The relation between both may differ in the individual cases: the difference will be much bigger in ethically and politically relevant questions than in the optimization of the marginal benefit of established technologies. Policy-advising TA only covers technology aspects which are subject to policy, like the safety and environmental standards, the protection of citizens against encroachment on their civil rights, the setting of priorities in research policy, the definition of framework conditions for innovations, etc. This is exactly where the largest part of policy-advising TA is taking place.

*Parliamentary* TA as a subcategory of policy-advising TA presupposes that parliaments play a crucial or at least an important and relevant role in technology governance: necessary assumption is *that parliamentary action is relevant for technology governance*. It is obvious that this assumption is facing problems since the role of parliaments in democratic decision processes is often categorized as declining, sometimes as hardly noticeable any more. The possibilities of parliamentary TA are limited not only by the restricted role of the state in technology governance but also by the restricted role of parliaments in the distribution of power in democratic systems. If TA is institutionalized in parliaments, its influence also depends on the respective institutional setting. In an analysis of the roles of parliamentary TA in technology governance based on a theory of institutions, a variety of possible combinations of different institutional configurations occurs (Cruz-Castro and Sanz-Menendez 2004), which is also enriched by the characteristics of the democratic institutions of a nation-state and various political traditions (Vig and Paschen 1999).[2]

[2]There is a lively and growing community of parliamentary TA in Europe which has organized itself in the European Parliamentary Technology Assessment (EPTA) Network.

## TA in Public Debate: Conflicts and Participation

Conflicts are characteristic of decisions in the field of technology, while consensus tends to constitute the exception. Making decisions in such conflict situations often results in *problems of legitimization* because there will be winners (who profit from specific decisions) and losers. This is frequently the case when decisions must be made about the site of a technical facility such as a nuclear power plant, a waste disposal plant, or a large chemical production plant. Depending on the selected location, people in the direct neighborhood will have to accept more disadvantages than others. Problems of legitimization always surface when the distribution of advantages and disadvantages is unequal.

In view of the decades of experience with a number of very serious acceptance problems and certain grave conflicts over technology, it has become clear that the question of legitimization is obviously important. Many examples can be given such as opposition to nuclear power, the problem of expanding airports, establishing new infrastructure elements such as highways or railway connections, the problem of how to dispose of radioactive waste, the release of genetically modified plants, the Strategic Defense Initiative ("Star Wars," SDI), and regional and local conflicts on waste disposal sites, waste incineration plants, and the location of chemical processing facilities. In these areas, political decisions are sometimes not accepted by those affected or by the general public, even though they are the result of democratic decision-making procedures. Conflict regulation and prevention are of the highest importance and a subject also to TA in its history.

Since the very beginnings of TA, there has been repeated demand for participative orientation, frequently following normative ideas from the fields of deliberative democracy or discourse ethics (Barber 1984; Renn and Webler 1998). According to these normative ideas, assessment and evaluation of technology should be left neither to the scientific experts (expertocracy) nor to the political deciders (decisionism) (see Habermas 1970 for this distinction). It is the task of participative TA to include societal groups – stakeholders, affected citizens, nonexperts, and the public in general – in assessing technology and its consequences. In this manner, participative TA procedures are deemed to improve the practical and political legitimacy of decisions on technology. Such TA is informed and advised by science and experts and, in addition, by people and groups external to science and politics (Joss and Belucci 2002).

The participation of citizens and of those affected is believed to improve the knowledge basis as well as the values fundament on which judgments are based and decisions are made. "Local" knowledge, with which experts and decision makers are often not familiar, is to be used in order to achieve the broadest possible knowledge base and to substantiate decisions. This discernibly applies especially to local and regional technological problems, in particular, to questions of siting. Furthermore, in a deliberative democracy, it is necessary to take the interests and values of ideally *all* those participating and affected into consideration in the decision-making process. Participation should make it possible for decisions on technology to be accepted by a larger spectrum of society

despite divergent normative convictions. In the end, this will also improve the robustness of such decisions and enhance their legitimacy (Gibbons et al. 1994). Several approaches and methods have been developed and applied in the recent years, such as consensus conferences, citizens' juries, and focus groups (Joss and Belucci 2002).

## TA for Shaping Technology

In the engineering sciences, the challenges with which TA is confronted have been discussed as demands on the profession of engineers. The value dimension of technology has been shown in many case studies, especially in engineering design processes (van Gorp 2005; van de Poel 2009). Decisions on technology design involve value judgments. In this respect there is, in other words, a close relationship between TA on the one side and professional engineering ethics and the ethics of technology on the other.

TA is one of a number of activities that provide orientation to and support for societal opinion forming and political decision making. Within the various approaches which can be subsumed under the social constructivist paradigm, the impact of those activities is primarily seen in the field of technology itself: ethical reflection aims to contribute to the technology paths, products, and systems to be developed (Yoshinaka et al. 2003). Theory-based approaches of *shaping technology* have been proposed, for example, by means of technology assessment (Rip et al. 1995) or variations of social construction of technology (Bijker and Law 1994). They have introduced strong claims for *influencing technology* by reflecting its social role and its consequences in the debate. The central message is that a "better" technology could be designed and constructed by using SST and CTA or other social constructivist approaches. The overall aim "to achieve better technology in a better society" (Schot and Rip 1997) shall be realized by looking at the very shape of technologies itself. The social construction of technology has even been extended to the social construction of the *consequences* of technology. In order to achieve a more environmentally and socially friendly technology, network-oriented approaches of the sociology of technology tried to control the problem of non-intended side effects of technology by applying adequate strategies of shaping technology during its genesis (Weyer et al. 1997).

## TA Approaches Relevant for Design for Values

Within the branch of TA addressing the shaping and design of technology directly (see section "TA for Shaping Technology"), several approaches have been developed of which some of them also have been implemented in practical projects. In this section I will briefly introduce the constructive TA, the approach of the Association of German Engineers, the Leitbild assessment, and ideas going for socially more compatible technology.

## Constructive Technology Assessment (CTA)

The basic assumption of CTA (which was developed in the Netherlands) is that TA meets with difficult problems of implementation and effectiveness whenever it concerns itself with the impacts of a technology after the latter has been developed or is even already in use (Rip et al. 1995). According to the control dilemma (Collingridge 1980), once the impacts are relatively well known, the chances of influencing them will significantly decrease because that knowledge will only be available in the later stages of development. It would therefore be more effective to accompany the process of the *development* of a technology constructively. The origin of technological impact is traced back to the development phase of a technology and the many decisions to be taken there so that dealing with the consequences of technology becomes a responsibility that already starts in the technology design phase. CTA argues for the early and broad participation of societal actors, including key economic players, users, and people affected in these early stages. In normative respect, CTA builds on a basis of deliberative democracy with a liberal picture of the state putting emphasis on self-organizing processes in the marketplace. Three processes have been proposed (Schot and Rip 1997, p. 257f.):

1. *Technology forcing*: Influencing technological progress through the promotion of research and technology as well as through regulation is how the state can intervene in technology (see section "TA for Policy Advice: The Case of Parliaments"). The influence of the state is, however, seen as restricted. CTA therefore also addresses other actors such as banks and insurance companies, standard bodies, and consumer organizations. Through their business and organizational policy, these institutions can directly intervene in certain technological innovations, for instance, by dispensing with chlorine chemistry, by investing in environmentally compatible manufacturing technology, or by developing social standards that are also valid for branches of a company located in developing nations.
2. *Strategic niche management*: Political promotion of technology and innovation should, according to CTA, be concerned with occupying "niches" in technology's repertory. In these niches publicly sponsored technology can – if protected by subsidies – be developed step by step, can make use of processes of learning, can gain acceptance, and finally can maintain its own in free competition unaided by public support (this part of CTA is related to the more policy-advising TA; see section "TA for Policy Advice: The Case of Parliaments").
3. *Societal dialogue on technology*: CTA regards it necessary to create the opportunities and structures for critical and open dialogue on technology. This dialogue must go beyond the limits of scientific discourse and expert workshops and, instead, include representatives from the economy and from the population. This postulate applies to technology forcing as well as to niche management.

CTA has been applied to a great variety and number of different technologies so that a huge body of experience exists (e.g., Rip et al. 1995; van Merkerk 2007).

## Leitbild Assessment

In Germany, the concept of empirical technology genesis research developed in parallel with CTA (Dierkes et al. 1992; Weyer et al. 1997). As in CTA, the paramount objective is to analyze the processes of shaping technology and of the embodiment of technology by society instead of looking on its impacts. The shaping and diffusion of technology are traced back to social processes of communication, networking, and decision making. TA accordingly consists of research into the social processes which contribute to technological design, analyzing the "setscrews" for intervening in these processes and informing decision makers on these findings. There is, in this concept, almost no further mention of technological impact; it is presumed that the unintended side effects could be completely or largely avoided by improving the process of technology shaping, in particular by involving the envisaged users, people possibly affected, and citizens with their particular views, perspectives, concerns, and values (Weyer et al. 1997).

The concept of a *Leitbild assessment* (Dierkes et al. 1992) was one of the attempts to draw more practical conclusions from that research. The empirical result of social sciences research was that technology development often follows broader and non-technological ideals which were called Leitbild (leitmotif, "guiding visions," cp. Grin and Grunwald 2000). Leitbilder are often phrased in metaphors which are shared, implicitly or explicitly, by the relevant actors. Some famous examples from earlier times of technology development and planning are the "paperless office," the "warfare without bloodshed," and the "automobile city." The expectation is that through societal construction of the Leitbilder, technology could be indirectly influenced in order to prevent any negative effects and to provide positive results. In a sense, this approach is a direct predecessor of Design for Values because social Leitbilder obviously include values – shaping technology according to a Leitbild therefore implies shaping technology considering particular values. However, the approach did not work in practice (Grunwald 2000b) because it was not made operable to be usable in the concrete workplaces where technology design takes place.

## The Concept of the Association of German Engineers

The Association of German Engineers (VDI, Verein Deutscher Ingenieure) considered challenges of technology to society from the 1960s. A lot of publications of VDI addressed issues such as technology and society, responsibility of engineers, and a code of conduct.

The most prominent outcome of these activities is the VDI guideline no. 3780 (VDI 1991, also available in English), which has become relatively widespread. It envisages a "Guide to Technology Assessment According to Individual and Social Ethical Aspects." For engineers and in industry, assessments are to a certain extent part of their daily work. Evaluations play a central role whenever, for instance, a line of technology is judged to be promising or to lead to a dead end; whenever the

chances for future products are assessed; whenever a choice between competing materials is made; or whenever a new production method is introduced to a company. Though evaluation may be commonplace in daily engineering practice, what is essentially new in this guideline for societal technological evaluation is its scope, which also includes the societally relevant dimensions of impacts as well as technical and economic factors. Technological evaluation should be conducted in line with societally acknowledged values. Eight central values forming the VDI "Value Octagon" have been identified: functional reliability, economic efficiency, prosperity, safety, health, environmental quality, personality development, and social quality (VDI 1991). These values are thought to influence technical action and fall under the premise (VDI 1991, p. 7): "It should be the objective of all technical action . . . to secure and to improve human possibilities in life."

The values identified by VDI shall be involved in processes of technology development, in particular in technology design. They shall virtually be *built into* the technology. Engineers or scientists should, on the basis of their knowledge and abilities, point the development of technology in the "right" direction by observing these values and avoiding undesirable developments. If this exceeds their authority or competence, engineers should take part in the corresponding procedures of technology evaluation. This mode of operation is rather close to Value Sensitive Design (cp. the respective chapter ▶ "Value Sensitive Design: Applications, Adaptations, and Critiques" in this volume) and to Design for Values. However, VDI did not put much attention on how to make this approach work. Therefore, the approach is well integrated in education of engineers at many technical universities but did not have much impact on concrete development yet.

## Shaping Technology Toward Social Compatibility

Specific concern was and still is given to declining acceptance of technology and increasing resistance, e.g., because of risk perception. Many studies on the acceptance of key technologies or technology in general have been conducted since the early 1980s. In some countries monitoring procedures have been established to observe any change of the level of acceptance in the population. The experience of technology conflicts since the 1970s around some key technologies like nuclear technology or the gene technology (which sometimes led to warlike scenarios in some countries) raised the question whether it would be possible to avoid such conflicts a priori. The idea behind this approach is that technology conflicts could be avoided by taking into account the presumed acceptance in technology development and design, i.e., by developing technology in accordance with the values, norms, and fears of people (von Alemann et al. (1992). A strict orientation of new technology to existing values and patterns of risk acceptance was expected to overcome acceptance problems. Within the proposed approaches to shape technology in this way to assume better social compatibility, the *stakeholders* of technology development (customers, citizens, political parties, authorities, social movements – all groups or persons affected by technology policy) shall be involved

in the decision-making process. The degree of involvement ranges from real participation in the decision-making processes to measuring the rates of acceptance by polls. The assumption is that if the people concerned are involved in the decision-making process, the result should find acceptance among them: "Technologies developed through such strategies will be socially more viable and accepted, which will enhance the economic viability of new products and processes" (Rip et al. 1995, p. 5).

This approach, however, which was proposed in the 1990s faces several difficulties. Among these is the philosophical criticism of a naturalistic fallacy being involved, the conclusion from factual acceptance to moral acceptability. Shaping technology in coincidence with the factual values of the majority of people does not guarantee that ethical standards will be met. Factual acceptance does not replace the necessity of ethical scrutiny and justification of issues under consideration. Furthermore, orientating technology policy directly to the currently accepted values of the majority of people runs into difficulties because of the possible lack of stability of factual acceptance. Shaping technology in accordance with the values of people at a certain time does not prevent the situation that the so-designed technology might become a problem or might even be rejected years later because of change of factual values, lifestyles, and behavioral patterns (Grunwald 2000b).

## Design for Values as a Specification of TA

There is no clear-cut definition of TA. In the contrary, TA usually is not *defined* but *explained* by pointing to the motivations and diagnoses behind, by addressing the expectations toward TA and by referring to its methodology and its institutional contexts (see the preceding sections of this chapter, also Grunwald 2009a). Following this story line a lot of manifestations of TA have been developed so far and are, partially, elements of current practice. These approaches are tailor-made for specific constellations such as parliamentary TA or citizen's participation.

In this broad understanding of TA, it is possible and seems to be adequate to consider Design for Value as a specific manifestation of TA. To be more specific, Design for Values can be considered as a specification of TA in its function of shaping technology mentioned above (section "TA for Shaping Technology"; see also section "TA Approaches Relevant for Design for Values"). It has in common with TA in general the rejection of any technology determinism and the idea that technology can be shaped, at least to a certain extent, according to social and ethical values as well as with regard to democratically determined objectives. The specific constellation of Design for Values may be characterized by (1) directly addressing the context of engineering and, in particular, engineering design (van de Poel 2009) and by (2) highlighting normative issues involved in the design of new technology.

Looking at the history and on the experiences of TA in general, the question might come up on what could be learned from TA which now has been operating for about four decades. Regarding the major motivation of Design for Values to contribute to the design of technology in a way that technology would be in a better

accordance with social and ethical values, some observations seemingly lead to a portion of skepticism. Those approaches in TA that are most relevant to Design for Values have not (yet) been very successful (cf. sections "Leitbild Assessment" and "The Concept of the Association of German Engineers"). Shaping technology with regard to social and ethical values seems to be a hard mission being confronted with obstacles and pitfalls. The experience of TA shows that the success of approaches such as Design for Values depends on several circumstances and boundary conditions.

A reference to the debate on shaping nanotechnology by taking into account ethical values might help to learn from that experience. In that field, highly ambitious models of social construction and *constructability* of technology were applied from the very beginning of an ethical debate on (Grunwald 2012). In the early time of the debate, ethical deliberation was expected to contribute directly to the development of nanotechnologies in order to achieve "better" nanotechnologies in the sense of being in better accordance with ethical values and societal goals. The following distinction on understanding the meaning of "formation of nanotechnologies" was proposed in order to better understand the consequences of the ethical debate on nanotechnology (Grunwald and Hocke-Bergler 2010):

- *Strong understanding*: "contribution to the formation of nanotechnologies" means "influencing the development of nanotechnology" in the sense of *directly* influencing the R&D agenda of nanosciences and, therefore, the further course of research and technology itself as well as its outcomes in terms of products and systems.
- *Weak understanding*: "formation of nanotechnologies" means "formation of the societal context of nanotechnologies," where the "context" could be the public perception, the positions of stakeholders, the interventions of regulators, etc. – with possible impacts on the embedding of technology into society and with a more indirect influence on nanotechnology at the level of products and systems.

The main result of a review of the developments of the past decades was that there is only weak evidence for the "strong" understanding of ethical contributions to the formation of nanotechnology. Ethical deliberation did not directly affect the nanosciences, but complemented the view on what should urgently be done in other fields of research (like nanotoxicology) or by motivating public debate and also contributed to nanotech as a public phenomenon. The main finding of Grunwald (2012, Chap. 10) is support for the weaker sense but rejection of the stronger one of "shaping nanotechnology by ethical reflection."

It would, however, be too early to draw the conclusion that Design for Values would not work because it apparently did not work in this particular field. The reason is the well-known control dilemma (Collingridge 1980). Nanotechnology at the beginning of the twenty-first century was in a much too early stage to be subject to Design for Values in a direct sense. Other cases show that taking ethical and social values into account in more concrete design processes can be an important element of the overall design of technology with positive results at the side of

products and systems (e.g., van Gorp 2005). Thus, it turns out that it is crucial to identify appropriate occasions in the research and innovation processes to influence the further design process by reflecting in values involved. These occasions will, following the nanotech example, presumably not be located in the very early stages of new and emerging sciences and technologies (NEST) but might be found in later stages where more concrete applications are addressed.

## References

Aichholzer G, Bora A, Bröchler S, Decker M, Latzer M (eds) (2010) Technology Governance. Der Beitrag der Technikfolgenabschätzung. Edition Sigma, Berlin

Barber BR (1984) Strong democracy. Participatory politics for a new age. University of California Press, Berkeley

Bechmann G (1994) Frühwarnung – die Achillesferse der TA? In: Grunwald A, Sax H (eds) Technikbeurteilung in der Raumfahrt. Anforderungen, Methoden, Wirkungen. Edition Sigma, Berlin, pp 88–100

Bechmann G, Decker M, Fiedeler U et al (2007) Technology assessment in a complex world. Int J Foresight Innov Policy 3:6–27

Beck U (1986) Risikogesellschaft. Auf dem Weg in eine andere Moderne. Suhrkamp, Frankfurt

Bijker WE, Law J (eds) (1994) Shaping technology/building society. MIT Press, Cambridge, MA

Bijker WE, Hughes TP, Pinch TJ (eds) (1987) The social construction of technological systems. New directions in the sociology and history of technological systems. MIT Press, Cambridge, MA

Bimber BA (1996) The politics of expertise in congress: the rise and fall of the office of technology assessment. State University of New York Press, New York

Collingridge D (1980) The social control of technology. Frances Pinter, New York

Cruz-Castro L, Sanz-Menendez L (2004) Politics and institutions: European parliamentary technology assessment. Technol Forecast Soc Change 27:79–96

Decker M, Ladikas M (eds) (2004) Bridges between science, society and policy. Technology assessment – methods and impacts. Springer, Berlin

Dierkes M, Hoffmann U, Marz L (1992) Leitbild und Technik. Zur Entstehung und Steuerung technischer Innovationen. Edition Sigma, Berlin

Durbin P, Lenk H (eds) (1987) Technology and responsibility. Reidel, Dordrecht

Gee D, Greenberg M (2002) Asbestos: from 'magic' to malevolent mineral. In: Harremoes P, Gee D, MacGarvin M, Stirling A, Keys J, Wynne B, Guedes Vaz S (eds) The precautionary principle in the 20th century. Late lessons from early warnings 1896–2000. Sage, London, pp 49–63

Gibbons M, Limoges C, Nowotny H, Schwartzmann S, Scott P, Trow M (1994) The new production of knowledge. Sage, London

Grin J, Grunwald A (eds) (2000) Vision assessment: shaping technology in 21st century society. Springer, Heidelberg

Grunwald A (2000a) Technik für die Gesellschaft von morgen. Möglichkeiten und Grenzen gesellschaftlicher Technikgestaltung. Campus, Frankfurt

Grunwald A (2000b) Technology policy between long-term planning requirements and short-ranged acceptance problems. New challenges for technology assessment. In: Grin J, Grunwald A (eds) Vision assessment: shaping technology in 21st century society. Springer, Heidelberg, pp 99–148

Grunwald A (2003) Technology assessment at the German Bundestag: 'expertising' democracy for 'democratising' expertise. Sci Pub Policy 30(3):193–198

Grunwald A (2006) Scientific independence as a constitutive part of parliamentary technology assessment. Sci Pub Policy 33(2):103–113
Grunwald A (2008) Nanoparticles: risk management and the precautionary principle. In: Jotterand F (ed) Emerging conceptual, ethical and policy issues in bionanotechnology. Springer, Berlin, pp 85–102
Grunwald A (2009a) Technology assessment: concepts and methods. In: Meijers A (ed) Philosophy of technology and engineering sciences (handbook of the philosophy of science), vol 9. Elsevier, Amsterdam, pp 1103–1146
Grunwald A (2009b) Vision assessment supporting the governance of knowledge – the case of futuristic nanotechnology. In: Bechmann G, Gorokhov V, Stehr N (eds) The social integration of science. Institutional and epistemological aspects of the transformation of knowledge in modern society. Edition Sigma, Berlin, pp 147–170
Grunwald A (2011) Energy futures: diversity and the need for assessment. Futures 43(8):820–830
Grunwald A (2012) Responsible nano(bio)technology: ethics and philosophy. Pan Stanford, Singapore
Grunwald A, Hocke-Bergler P (2010) The risk debate on nanoparticles: contribution to a normalisation of the science/society relationship? In: Kaiser M, Kurath M, Maasen S, Rehmann-Sutter C (eds) Governing future technologies. Nanotechnology and the rise of an assessment regime. Springer, Dordrecht, pp 157–177
Habermas J (1970) Toward a rational society: student protest, science, and politics. Beacon Press, Boston. First publication: Habermas J (1968) (ed) Technik und Wissenschaft als Ideologie. Suhrkamp, Frankfurt
Habermas J (2001) Die Zukunft der menschlichen Natur. Auf dem Weg zur liberalen Eugenetik? Suhrkamp, Frankfurt. English version: The future of the human nature (trans: Hella Beister and William Rehg). Polity Press, Cambridge
Harremoes P, Gee D, MacGarvin M, Stirling A, Keys J, Wynne B, Guedes Vaz S (eds) (2002) The precautionary principle in the 20th century. Late lessons from early warnings. Sage, London
Joss S, Belucci S (eds) (2002) Participatory technology assessment – European perspectives. Westminster University Press, London
Paschen H, Petermann T (1992) Technikfolgenabschätzung – ein strategisches Rahmenkonzept für die Analyse und Bewertung von Technikfolgen. In: Petermann T (ed) Technikfolgen-Abschätzung als Technikforschung und Politikberatung. Campus, Frankfurt, pp 19–42
Renn O, Webler T (1998) Der kooperative Diskurs – Theoretische Grundlagen, Anforderungen, Möglichkeiten. In: Renn O, Kastenholz H, Schild P, Wilhelm U (eds) Abfallpolitik im kooperativen Diskurs – Umweltplanung im kooperativen Diskurs. Bürgerbeteiligung bei der Standortsuche für eine Deponie im Kanton Aargau. vdf Hochschulverlag, Zürich, pp 3–103
Rip A, Misa T, Schot J (eds) (1995) Managing technology in society: the approach of constructive technology assessment. Pinter Publishers, London
Schot J, Rip A (1997) The past and future of constructive technology assessment. Technol Forecast Soc Change 54(2–3):251–268
Siune K, Markus E, Calloni M, Felt U, Gorski A, Grunwald A, Rip A, de Semir V, Wyatt S (2009) Challenging futures of science in society. Report of the MASIS expert group. European Commission, Brussels
Smits R, den Hertog P (2007) TA and the management of innovation in economy and society. Int J Foresight Innov Policy 3(1):28–52
van de Poel I (2009) Values in engineering design. In: Meijers A (ed) Philosophy of technology and engineering sciences, vol 9. North Holland, Amsterdam, pp 973–1006
van Gorp A (2005) Ethical issues in engineering design, safety and sustainability, vol 2, Simon Stevin series in the philosophy of technology. 3TU Ethics, Delft
van Merkerk R (2007) Intervening in emerging technologies – A CTA of lab-on-a-chip technology. Utrecht University, Royal Dutch Geographical Society, Utrecht
Verein Deutscher Ingenieure (1991) Technikbewertung – Begriffe und Grundlagen: Erläuterungen und Hinweise zur VDI-Richtlinie 3780, vol 15, VDI Report. VDI, Düsseldorf

Vig N, Paschen H (eds) (1999) Parliaments and technology assessment. The development of technology assessment in Europe. State University of New York Press, Albany

von Alemann U, Schatz H, Simonis G (1992) Leitbilder Sozialverträglicher Technikgestaltung. VS Verlag für Sozialwissenschaften, Opladen

von Schomberg R (ed) (1999) Democratizing technology. Theory and practice of a deliberative technology policy. International Centre for Human and Public Affairs, Hengelo

von Schomberg R (2005) The precautionary principle and its normative challenges. In: Fisher E, Jones J, von Schomberg R (eds) The precautionary principle and public policy decision making. Edward Elgar Publishing, Cheltenham, pp 141–165

Weyer J, Kirchner U, Riedl L, Schmidt JFK (1997) Technik, die Gesellschaft schafft. Soziale Netzwerke als Ort der Technikgenese. Edition Sigma, Berlin

Yoshinaka Y, Clausen C, Hansen A (2003) The social shaping of technology: a new space for politics? In: Grunwald A (ed) Technikgestaltung: zwischen Wunsch oder Wirklichkeit. Springer, Berlin, pp 117–131

# Part II

# Perspectives

# Conflicting Values in Design for Values

Ibo van de Poel

## Contents



**Abstract**

Designers are regularly confronted with conflicting values in design: different values select different design options as best. This contribution deals with how one can deal with such value conflicts in design for values. A characterization of value conflict in design is given, and the notion is compared with the notion of moral dilemmas. It is further argued that value conflicts in design entail a kind of multi-

I. van de Poel (✉)
Department of Ethics and Philosophy of Technology, Delft University of Technology, Delft, The Netherlands
e-mail: i.r.vandepoel@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_5

criteria decision problems to which Arrow's impossibility theorem applies. This theorem says that there is no procedure to aggregate scores on individual criteria (values) into an overall score unless one is willing to violate one of more minimally reasonable conditions for any such an aggregation procedure. Six methods to deal with value conflicts (cost-benefit analysis, direct trade-offs, maximin, satisficing, judgment, and innovation) are discussed. Three of these avoid Arrow's theorem by assuming a form of value commensurability, although they may be too informationally demanding and have other disadvantages as well. The other three are non-optimizing methods that do not result in one best solution and therefore do not entirely solve the value conflict, although they are a way forward in some respects. In conclusion, an approach that combines the several methods is proposed as a way to deal with cases of conflicting *moral* values in design and which avoids many of the disadvantages of the single methods.

**Keywords**

Design for values • Value conflict • Arrow's impossibility theorem • Value commensurability • Trade-offs • Multi-criteria problems • Moral dilemmas • Value sensitive design

## Introduction

In design, we often try to respect or promote a range of values such as safety, sustainability, human welfare, privacy, inclusiveness, and justice. When we design not for one value but for a range of values, it will regularly occur that design options that score good on one value score less on another. In other words, if we use the values as choice criteria, for example, during concept selection, one value will point in the direction of one particular design and another value in the direction of another. How can we deal with such value conflicts in design?

There are lots of methods in traditional engineering to deal with trade-offs and conflicts between evaluation criteria which are in principle also relevant for value conflicts. This includes such methods as multi-criteria design analysis, the method of weighted objectives, Pugh charts, and quality function deployment (QFD). Most of the methods, however, do not pay explicit attention to their value dimension. Moreover, it has been argued that many of these methods are methodologically flawed. Franssen (2005) has shown that Arrow's impossibility theorem also applies to multi-criteria decisions in engineering design. Applied to design, this theorem implies that it is impossible to aggregate scores of options on individual criteria into an overall ordering of the options on all criteria without violating one or more axioms, which any reasonable aggregation procedure should minimally meet.

Value conflict, then, is a persistent problem in engineering design. The aim of this chapter is to explore ways in which we can deal with value conflict if one designs for values. I start with further exploring the notion of value conflict. In doing so, I will also relate value conflict to moral dilemmas and will provide a number of examples of value conflict in engineering. In the next section, I explain

how Arrow's impossibility theorem applies to multi-criteria decision problems in design, and I discuss its consequences and conditions under which it may be avoided. In section "Approaches for Dealing with Value Conflict," I explore six different methods for dealing with value conflicts in design. In the final concluding section, I compare the methods and propose an approach that combines different methods to deal with conflicts between moral values in design.

## Value Conflict in Engineering Design

### Value Conflict and Moral Dilemmas

A value conflict may be defined as the situation in which all of the following conditions apply (Van de Poel and Royakkers 2011):

1. A choice has to be made between at least two options for which at least two values are relevant as choice criteria.
2. At least two different values select at least two different options as best.
3. There is not one value that trumps all others as choice criterion. If one value trumps another, any (small) amount of the first value is worth more than any (large) amount of the second value.

The reason for the second condition is that if all values select the same option as the best one, we can simply choose that one, so that we do not really face a value conflict. The reason for the third condition is that if one value trumps all others, we can simply order the options with respect to the most important value, and if two options score the same on this value, we will examine the scores with respect to the second, less important, value, and so on. So if values trump each other, there is not a real value conflict.

Value conflicts are somewhat similar, though not entirely the same as moral dilemmas. Williams (1973, p. 180) provides the following general characterization of moral dilemmas:

1. The agent ought to do a.
2. The agent ought to do b.
3. The agent cannot do a and b.

Moral dilemmas are thus formulated in terms of conflicting "oughts" which cannot be followed at the same time. It is instructive to try to formulate value conflicts in terms of moral dilemmas to see the similarities and differences between both types of conflict. On the basis of the earlier conditions, we might give the following characterization of a value conflict for the case of two values (v and w) and two options (a and b):

1. Value v selects option a as best.
2. Value w selects option b as best.

3. The values v and w do not trump each other.
4. It is impossible to choose both a and b.

This formulation would amount to a moral dilemma if the following two conditions are also met:

5. Option a ought to be chosen because value v selects it as best.
6. Option b ought to be chosen because value w selects it as best.

Statements (5) and (6) are, however, far from uncontroversial. There are in fact two independent objections possible against (5) and (6). The first objection is that if in a choice situation two values v and w are relevant, an "ought" cannot follow from considering only one of the values unless that value trumps the other (which is in fact denied by (3)). The reason for this is that "ought" judgments are all things considered judgments that take into account all relevant considerations in the (choice) situation.[1] A second possible objection is that (5) and (6) seem to presuppose that we ought to choose what brings about most value. This maximizing assumption is indeed present in many choice methodologies and in some ethical theories. It is, however, not an uncontroversial assumption.[2] Therefore, I think it is better to avoid this assumption in a characterization of value conflict.

If a value conflict is indeed characterized by (1)–(4), a value conflict does not entail a moral dilemma, although value conflict may occasionally amount to moral dilemmas. But even if value conflicts do not necessarily include hard choices in which two (or more) "oughts" conflict, they still involve difficult choices, and it may be hard to know not only what to choose but also how to choose. In the remainder of the chapter, I will focus on value conflicts but I will come back to moral dilemmas in the conclusion.

## Examples of Value Conflict Engineering Design

Let me now turn to a number of examples of value conflict in engineering design. This will both make the above discussion more tangible and, at the same time, it will show how this is relevant for design for values.

**Example 1: Safety Belts** Suppose you are a car designer and you are designing a safety seat belt system for a car. You know that the use of seat belts in cars reduces

[1]This is exactly why moral dilemmas have been characterized as hard choices (Levi 1986) and, also, why some philosophers deny the possibility of moral dilemmas because according to them there is always an overall point of view in which only one ought applies.

[2]What makes the assumption problematic is not just the assumption that "more value is better than less" but also the assumption that there is an obligation (an "ought") rather than just a recommendation to choose the highest value. Even if "more value is better than less," we might not be obliged to choose the object with higher value.

**Table 1** Different seat belt systems for cars

| | Safety | Freedom |
|---|---|---|
| Traditional seat belt | Lowest | Highest |
| Seat belt with warning signal | Moderate | Moderate |
| Automatic seat belt | Highest | Lowest |

the number of fatalities and injuries, that is, a car driver or other car occupants have a lower probability to get killed or injured (or they are less severely injured) in case of a car accident if they wear a seat belt than if they do not. You also know, however, that some people tend to forget the use of seat belts or find them unpleasant to use or do not use them for other reasons.

Two values that are relevant in the design of a seat belt system then are safety and freedom. Safety is here mainly understood as lower probability of fatality or injury, or less severe injuries, in case of an accident for car drivers or occupants. Freedom is by and large understood as the presence of a free and uninfluenced choice in using a safety belt or not.

Let us suppose that there are three options that you have selected to choose between as the designer of a safety belt system: (1) a traditional seat belt; (2) a so-called automatic seat belt that enforces its use, for example, by making it impossible to enter the car without using the seat belt or making it impossible to start and drive the car without using the seat belt; and (3) a system with a warning signal that makes an irritating noise if the seat belt is not used.

Table 1 represents these three options and their scores on the values of safety and freedom. The choice situation is an example of a value conflict as I defined it above. The question is how the designer who wants to design a seat belt for both the value of safety and the value of freedom can choose between the three options.

**Example 2: The Storm Surge Barrier in the Eastern Scheldt** After a huge flood disaster in 1953, in which a large number of dikes in the province of Zeeland, the Netherlands, gave way and more than 1,800 people were killed, the Delta plan was drawn up.[3] Part of this Delta plan was to close off the Eastern Scheldt, an estuary in the southwest of the Netherlands. From the end of the 1960s, however, there was growing societal opposition to closing off the Eastern Scheldt. Environmentalists, who feared the loss of an ecologically valuable area because of the desalination of the Eastern Scheldt and the lack of tides, started to resist its closure. Fishermen also were opposed to its closure because of the negative consequences for the fishing industry. As an alternative, they suggested raising the dikes around the Eastern Scheldt to sufficiently guarantee the safety of the area.

In June 1972, a group of students launched an alternative plan for the closure of the Eastern Scheldt. It was a plan that had been worked out as a study assignment by students of the School of Civil Engineering and the School of Architecture of the Technical University of Delft and the School of Landscape Architecture of the

[3]A more detailed discussion can be found in van de Poel (1998).

**Table 2** Options for Eastern Scheldt

| | Safety | Ecology |
|---|---|---|
| Closing off Eastern Scheldt | Highest | Lowest |
| Storm surge barrier | Moderate | Moderate |
| Raising dikes around the Eastern Scheldt | Lowest | Highest |

Agricultural University of Wageningen. The values the students focused on were safety and ecological care. On the basis of these values, they proposed a storm surge barrier, i.e., a barrier that would normally be open and allow water to pass through but that could be closed if a flood threatened the hinterland.

Table 2 lists the three abovementioned options. The original plan to close off the Easter Scheldt would be the safest (in terms of probability of flooding and number of fatalities in case of flooding) but scores the worst in terms of ecology. Heightening the dikes would most likely be the least safe (although this was not entirely beyond debate) and the best in terms of ecology. The storm surge barrier was a creative compromise between both values.

**Example 3: Refrigerants for Household Refrigerators** As a consequence of the ban on CFCs in the 1990s, an alternative to CFC 12 as refrigerant in household refrigerators had to be found.[4] I will focus here on three (moral) values that played an explicit role in the search for alternative coolants: safety, health, and environmental sustainability. In the design process, safety was mainly understood as nonflammability, and health as nontoxicity. Both understandings were based on existing codes, standards, and testing procedures like the ASHRAE Safety Code for Mechanical Refrigeration. ASHRAE is the American Society of Heating, Refrigerating, and Air-Conditioning Engineers. Environmental sustainability was typically formulated in terms of a low ODP (ozone depletion potential) and a low GWP (global warming potential). Both ODP and GDP mainly depend on the atmospheric lifetime of refrigerants. In the design process, a conflict between those three considerations arose. This value conflict can be illustrated with the help of Fig. 1.

Figure 1 is a graphic representation of CFCs based on a particular hydrocarbon. In the top, there is methane or ethane or another hydrocarbon. If one moves to the bottom, hydrogen atoms are replaced by either chlorine atoms (if one goes to the left) or fluorine atoms (if one goes to the right). In this way, all CFCs based on a particular hydrocarbon are represented. The figure shows how the properties flammability (safety), toxicity (health), and atmospheric lifetime (sustainability) depend on the exact composition of a CFC. As can be seen, minimizing the atmospheric lifetime of refrigerants means maximizing the number of hydrogen atoms, which increases flammability. This means that there is a fundamental trade-off between

[4]A more detailed discussion can be found in van de Poel (2001). The data for GWP given in Table 3 are based on a more recent IPCC report (Solomon et al. 2007, Table 2.14) and therefore deviate from the data in van de Poel (2001). The GWP for a 100-year time horizon is given in the table.

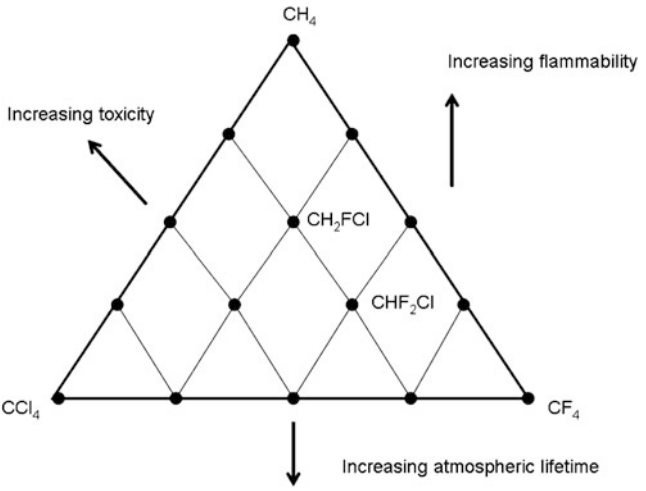



**Fig. 1** Properties of refrigerants (Based on McLinden and Didion (1987))

**Table 3** Properties of refrigerants

| | Environmental sustainability | | Health | Safety |
|---|---|---|---|---|
| | ODP | GWP | Toxicity class | Flammability class |
| CFC 12 | 1 | 10,900 | A | 1 |
| HFC 134a | 0 | 1,430 | A | 1 |
| HFC 152a | 0 | 124 | A | 2 |
| HC 290 (propane) | 0 | 3 | A | 3 |
| HC 600a (isobutane) | 0 | 3 | A | 3 |

flammability and environmental effects, or between the values of safety and sustainability.

Table 3 lists some of the options that were considered as replacements for CFC 12 as coolant in household refrigerators. The ODP (ozone depletion potential) is measured relative to CFC 12, the global warming potential (GWP) relative to $CO_2$. For health, two toxicity classes have been defined in relevant codes and standards; class A is considered toxic and class B nontoxic. The same codes and standards define three flammability classes; class 1 is considered nonflammable, class 3 highly flammable, and class 2 moderately flammable.[5] The coolants listed in Table 3 exemplify a value conflict specifically between the value safety and environmental sustainability.

[5]It should be noted that coolants in the same flammability class are not necessarily equally flammable; neither are coolants in the same toxicity class necessarily equally toxic. Membership of certain flammability or toxicity class is determined by certain threshold values, and therefore does not reflect degrees of flammability or toxicity that follow the patterns shown in Fig. 1.

## Arrow's Theorem and Multi-criteria Decision-Making

It has been shown that the famous Arrow's theorem from social choice (Arrow 1950) also applies to multi-criteria decision-making (May 1954; Arrow and Raynaud 1986; Franssen 2005). Since value conflicts are a kind of multi-criteria decision problems, it also applies to value conflicts. Arrow's theorem establishes the impossibility of certain solutions that meet a number of minimally desirable characteristics. It thus also sets serious limits to ways to deal with value conflicts. This section discusses these limitations and how they follow from Arrow's theorem. The next section will, then, on the basis of this information, discuss approaches to deal with value conflicts.

### The Decision Problem for the Designer

Let me begin with formulating the decision problem that a designer or a design team faces when a value conflict occurs in design. I will assume here that there is one designer who faces a value conflict and makes a decision. This is of course a simplification compared to reality where often a design team is involved and where the additional complexity faced is how to reach a decision together.[6] I will further assume that the designer aims to design for values. So in dealing with the value conflict, the designer does not aim at the design that meets his/her own personal preferences best but rather he/she looks for a design that best meet the relevant values at stake. The designer may be said to take an ethical point of view.[7]

The decision problem faced by the designer may now be modeled as follows:

1. In the choice situation S, n values $v_1 \ldots v_i \ldots v_n$ are relevant.
2. In the choice situation S, m options $o_1 \ldots o_j \ldots o_m$ are feasible.
3. For each value $v_i$, a corresponding ordinal value function exists so that $v_i(o_a) \geq v_i(o_b)$ implies that option $o_a$ is at least as good (or better) as option $o_b$ with respect to value $v_i$.

In lay terms, (3) says that it is possible to order the options for each relevant value on a scale from better to worse. This implies at least an ordinal measurement of the options on each of the relevant values. Table 4 explains the difference between different measurement scales.

Below, we will consider the question whether it is possible to derive on the basis of the information contained in (1)–(3) a value function $v(o_j)$ that orders all options

[6]To this social choice problem, the Arrow's theorem also applies (Kroes et al. 2009).

[7]According to some ethical theories (e.g., Kantianism), all people would come to the same conclusion if they take an ethical point of view. In that case, the earlier assumption of there being just one decision-maker would be justified.

**Table 4** Measurement scales

| Measurement scale | Invariant | Allowed transformation | Degree of freedom | Example |
|---|---|---|---|---|
| Ordinal | Order | Monotonic | Infinite | Mohs scale of mineral hardness[a] |
| Interval | Ratio between differences | Positive linear | Two (zero point and unit) | Temperature measured in Celsius |
| Ratio | Ratio | Scalar | One (unit) | Distance measured in meters |

[a]On this scale, the hardness of a material is measured by determining the hardest material that the given material can scratch and/or the softest material that can scratch the given material

on (at least) an ordinal scale with respect to all values $v_1 \ldots v_i \ldots v_n$ *in combination*. Since we are looking for the general possibility to construct such a value function from (1)–(3), this includes cases when the designer faces a value conflict.

## Arrow's Theorem

Let me now turn to Arrow's theorem. Arrow considered the following social choice situations:

1. In the social choice situation S, n individuals $w_1 \ldots w_i \ldots w_n$ are involved in the decision.
2. In the choice situation S, m options $o_1 \ldots o_j \ldots o_m$ are feasible.
3. For each individual $w_i$, a corresponding ordinal value function exists so that $w_i(o_a) \geq w_i(o_b)$ implies that option $o_a$ is at least as good (or better) as option $o_b$ in terms of the preferences of individual $w_i$.

Again, (3) says that each individual can order all options on a scale from best to worst (allowing for indifferences between some options).

Arrow shows that if $n \geq 2$ and $m \geq 3$, it is impossible to find a function or decision procedure that meets a number of minimally reasonable conditions to translate the individual preferences into a collective preference. These minimal conditions are[8]:

- *Collective rationality*. This condition implies that the collective preference ordering must be complete and transitive. A preference ordering is complete if all alternatives are ordered by it. Transitivity requires that if $o_a$ is ordered over $o_b$ and $o_b$ is ordered over $o_c$, $o_a$ is also ordered over $o_c$.

[8]The requirements given are somewhat weaker than those originally formulated by Arrow. See, e.g., Sen (1970). See also Franssen (2005).

- *Unrestricted domain*. This condition implies that there are no restrictions with respect to how an individual orders the alternatives, apart from conditions of completeness and transitivity for the individual preference orderings.
- *Pareto principle*. This condition implies that if everyone prefers $o_a$ over $o_b$, the collective preference ordering should order $o_a$ over $o_b$.
- *Independence of irrelevant alternatives*. The ordering of alternative $o_a$ relative to alternative $o_b$ may not depend on the inclusion or exclusion of a third alternative in the set of alternatives.
- *Absence of a dictator*. This condition implies that there is no individual whose preferences determine the collective preference.

Arrow's theorem means that no *general* procedure exists to translate individual preferences into a collective preference ordering unless one is willing to breach one of the abovementioned conditions.

## Application of Arrow's Theorem to Multi-Criteria Decision Problems

It can easily be seen that the choice situation faced by the designer described above is structurally similar to the social choice situation to which Arrow's theorem applies. The difference is that where in the original Arrow case, individuals order the alternatives, in the design choice situation, values order the alternatives.[9] In both cases, the possibility of an aggregation procedure that meets some minimally desirable characteristics is at stake.

Franssen (2005) has argued that Arrow's theorem also applies to multi-criteria choices, and he also argued that all conditions listed above that play a role in Arrow's theorem are still reasonable in the case of multi-criteria decision problems in engineering (see also Jacobs et al. 2014). I will not repeat all of his arguments but will focus on some of the main issues with respect to the applicability of Arrow's conditions to choices with various values in engineering design.

One possible objection against the requirement of *collective rationality* is that in design, we only want to select the best design and we have no interest in ordering all the other designs. It can, however, be shown that if we release the condition accordingly, an impossibility theorem that is rather similar to the original Arrow's theorem can be proven (Mas-Colell and Sonnenschein 1972).

With respect to the condition of *unrestricted domain*, one might argue that *given* a specific value and a specific range of options, the ordering of those options on that value is not unrestricted. We cannot suddenly say in the safety belt example (example 1 in section "Value Conflict in Engineering Design") that the traditional

[9] A difference is that the values have different degrees of importance, whereas in the original Arrow choice situation, each individual has equal weight. We can, however, repair this by replacing each value by x values where x is the (relative) degree of importance of that value (cf. Franssen 2005).

seat belt is best in terms of safety. The condition of unrestricted domain is, however, to be understood as expressing that we look for a procedure that is able to deal with any way, a value may order the alternatives (as long as the conditions of completeness and transitivity are met). The Arrow's theorem thus shows the impossibility of a generally applicable procedure, which does not imply the impossibility of solving one particular case.

The *Pareto principle* says that if all values select an option as best that option should be ordered as best in the overall ordering. This seems hardly contestable. Still, there are two possible objections. One is that more value is not always better; sometimes we want to minimize a value (or a criterion for a value), or sometimes we might strive for a target rather than for as much as possible. However, such cases can usually be mathematically converted into a new criterion in which more is indeed better. A second objection is that sometimes the desirable degree of attainment of one value may be dependent on the actual attainment of another value. For example, suppose that two values in the design of a car are "safety" and "looks robust." It might be that for a not so safe car, one prefers a less robust looking design over a more robust looking design, while for the safer car, one prefers the more robust looking design over the less robust looking design (e.g., because one believes that car look should represent the underlying safety features). In cases like this, the Pareto principle does not apply.

The condition of *independence of irrelevant alternatives* seems a quite reasonable condition again. Also, in design, one does not want the choice between two alternatives to depend on the inclusion of a third in the overall choice. Underlying this condition, there are, however, two assumptions in the case of collective choice that Arrow made that have been contested. One assumption is that individual preferences can only be measured on an ordinal scale (and not on an interval or ratio scale). The other is the assumption of the impossibility of interpersonal utility comparison: we cannot compare the utility (also not on an ordinal scale) of one person with that of another.

It has been shown that if the first assumption is somewhat released and we allow preference or utility measurement for individuals on an interval scale, impossibility theorems similar to that of Arrow can be formulated (Hylland 1980). However, it has also been shown that under stronger assumptions about the informational base, aggregation procedures that meet axioms comparable to the ones proposed by Arrow are possible (Roberts 1980; Jacobs et al. 2014). If these assumptions are translated to the context of engineering design, aggregation would be possible in each of the following cases:

1. The score of all options on all individual values (criteria) can be measured on a ratio scale with respect to preference (utility) or value[10] (ratio measurability).

[10]If we assume only nonnegative utilities, ratio scale measurement is enough. However, if we also allow negative utilities, an additional commensurability assumption is needed for a reasonable aggregation procedure to be available (Tsui and Weymark 1997).

2. The score of all options on all individual values (criteria) can be measured on interval scales which share a common unit of measurement (unit commensurability).
3. The score of all options on all individual values (criteria) can be measured on a common ordinal scale, so that the score of the $^{xth}$ option on the $^{ith}$ criterion can be ordinally compared with the score of the $^{yth}$ option on the $^{jth}$ criterion (level commensurability).

If any of these three conditions apply, Arrow's theorem can be avoided. It should be noted that while the second and third conditions list a commensurability condition, the first condition does not require commensurability. It only requires that each individual value can be measured on a ratio scale, but this needs be a common scale, so that no commensurability is required. It should further be noted that unit commensurability and level commensurability are independent from each other; you can have unit commensurability without level commensurability, and vice versa. In the remainder I will speak of *value commensurability*, if either unit or level commensurability applies, or if both apply. If both apply, this is also sometimes called full commensurability. I will call two values incommensurable if neither unit nor level commensurability applies.

Franssen (2005) argues that ratio scale measurements of preferences or value (the first condition above) are impossible since ratio scales require extensive measurement, which he believes to be impossible for mental constructs like preference or value. Nevertheless, an often used approach, cost-benefit analysis, may be said to be based on the assumption that money can measure utility or value on a ratio scale (although this assumption is by no means unproblematic). I will discuss cost-benefit analysis in the next section as a possible method to deal with value conflicts; there, I will also discuss two methods that are available if one assumes either unit commensurability (direct trade-offs) or level commensurability (maximin).

The condition *absence of a dictator* in the case of multi-criteria problems implies that there is not one criterion or value that dictates the overall ordering of options. The criterion is the same as the third criterion that I formulated for value conflicts that there is no one value that trumps all others as choice criterion.

## Approaches for Dealing with Value Conflict

In this section, I will discuss the main approaches to value conflict and their advantages and disadvantages. The methods that will be discussed are:

- Cost-benefit analysis
- Direct trade-offs
- Maximin
- Satisficing (thresholds)
- Respecification
- Innovation

The first three methods each suppose a specific form of value commensurability through which Arrow's theorem might be avoided as we have seen in the previous section.[11] The other three methods are so-called non-optimizing methods (Van de Poel 2009).[12] They do not aim for one best option, and they do not, or at least not always or necessarily, result in one option that is to be chosen. They therefore do not meet Arrow's condition of collective rationality. Still, they may be useful in dealing with value conflicts in certain circumstances as will become clear.

## Cost-Benefit Analysis

In cost-benefit analysis, all relevant considerations are expressed in one common monetary unit, like dollars or euros. Because all values are measured on a common ratio scale (money), cost-benefit analysis assumes both ratio measurability and value commensurability. The advantage of this assumption is that Arrow's theorem is avoided and that it becomes possible to select the best alternative by expressing the score of options on a range of values in a common measure: money.

If we want to apply cost-benefit analysis to value conflicts in engineering design, we somehow need to express gains and losses in values, like freedom, safety, sustainability, etc., in monetary terms. A glance at the examples in section "Value Conflict in Engineering Design" shows how difficult this is. Take the safety belt example: is there a way to express the different degrees of freedom and safety realized by the various designs in monetary terms, and if so, can it be done in a reliable and uncontroversial way? If we look at the second example (the Eastern Scheldt barrier), a cost-benefit analysis was done for the original Delta plan, which still assumed a closed barrier in the Eastern Scheldt (Tinbergen 1959). In this cost-benefit analysis, safety was treated as an imponderable value, i.e., as a value that cannot be expressed in monetary terms.[13] However, ecology and environmental

[11]Cost-benefit analysis does not only suppose ratio measurability but also value commensurability because the various values are measured on a common scale (i.e., money). Theoretically, it would be possible to construct a method that only supposes ratio measurability (but see also note 10). Such a method could, for example, proceed by multiplying the score on one value dimension (measured on a ratio scale) with the score on another value dimension (measured on another ratio scale) and then selecting the alternative with the highest score so attained. It seems, however, doubtful whether that leads to a meaningful decision procedure for design, and as far as I know, no such decision methods have been proposed in the design literature. The multiplication of, for example, safety and sustainability, both measured on a (positive) ratio scale, for example, does not seem meaningful as a decision criterion. Note that in physics, such multiplications are sometimes meaningful, e.g., mass times velocity is a measure of momentum.

[12]The three non-optimizing methods discussed are the same as in Van de Poel (2009). There, I also discuss a fourth (diversity) that is not discussed here as it cannot be employed in a single design process (which I take to be the focus of this chapter). The discussion follows Van de Poel (2009) but has been updated and somewhat revised in several respects.

[13]The resulting costs can then be interpreted as the amount of money that one should be willing to pay for the increase in safety that is obtained by realizing the Delta plan.

concerns were not taken into account in the original cost-benefit analysis. It might be argued that these values are also imponderable. However, if one treats both the (conflicting) values of safety and ecology as imponderable, a cost-benefit analysis is of no help in example 2.

Despite the above reservations, approaches and methods like contingent validation have been developed to express considerations like safety, freedom, and ecology in monetary terms. Contingent validation proceeds by asking people how much they are willing to pay for a certain level of safety or for, for example, the preservation of a piece of beautiful nature. In this way, a monetary price for certain safety levels or a piece of nature is determined. Such methods are, however, beset with methodological problems, and it is questionable whether they deliver a reliable measurement for the values at stake. For example, the monetary value of a piece of nature is lower if one asks people how much they are willing to pay for it than if one asks for how much one would want to be compensated for giving it up (Horowitz and McConnell 2002). It has been suggested that such differences may be due to the intrinsic (moral) value of nature (Boyce et al. 1992).

There are a number of more fundamental issues with cost-benefit analysis as well. For one thing, it is questionable whether one could regard money as a good measure for preference or utility (as is assumed, as we saw in section "Arrow's Theorem and Multi-Criteria Decision-Making," if one conceives of money as a way to measure utility on a ratio scale). One problem here is the diminished marginal utility of money. For most people, a gain in income from 100 to 200 euros will imply a larger increase in utility than a gain in income from 10.100 to 10.200 euros, while both increases in utility should be the same if money is to measure utility on a ratio scale. Another problem is that it is questionable whether a similar gain in income, say 100 euros, will realize the same increase in utility for two different persons.

Another fundamental problem is whether we can measure a range of values like safety, sustainability, freedom, justice, etc., in terms of a common measure on a ratio scale (be it in terms of money, utility, or whatever other value measure). This is not just a practical or methodological issue about how to express these values in monetary terms (as discussed above), but it involves a more fundamental assumption about the nature of values. It should be noted that if one assumes that values are commensurable on a ratio scale, a loss in one value can always be compensated by a gain in another value (if the latter gain is large enough). Some authors believe that this assumption is wrong for at least some values. Consider, for example, the following trade-off: for how much money are you willing to betray your friend? It may well be argued that accepting a trade-off between friendship and financial gain undermines the value of friendship. On this basis, it is constitutive of the value of friendship to reject the trade-off between friendship and financial gain (Raz 1986). Such constitutive incommensurability seems especially true of moral values and values that regulate the relations between, and the identities of, people.

Even if some of the above issues are solved (or are just neglected as is often the case in actual cost-benefit analyses), one faces a range of additional methodological

and ethical issues in cost-benefit analysis (Hansson 2007). One issue is how to discount future benefits from current costs (or vice versa). One dollar now is worth more than one dollar in 20 years, not only because of inflation but also because a dollar now could be invested and would then yield a certain interest rate. To correct this, a discount rate is chosen in cost-benefit analysis. The choice of discount rate may have a major impact on the outcome of the analysis. Another issue is that one might employ different choice criteria once the cost-benefit analysis has been carried out. Sometimes all of the options in which the benefits are larger than the costs are considered to be acceptable. However, one can also choose the option in which the net benefits are highest or the option in which the net benefits are highest as a percentage of the total costs.

From the above considerations and reservations, it does *not* follow that one should never use cost-benefit analysis to deal with value conflicts in design. As we will see below, other approaches for dealing with value conflicts have their problems and disadvantages as well. In some design decision contexts, the above concerns may be less serious or we might have reasons to prefer cost-benefit analysis over other approaches. Still, one should be aware of the abovementioned limitations and issues in applying cost-benefit analysis.

## Direct Trade-Offs

A second approach to deal with value conflict is to make direct trade-offs between the relevant values. As we have seen in section "Arrow's Theorem and Multi-Criteria Decision-Making," this requires that the individual values are measured on (at least) an interval scale and that there is unit commensurability between the relevant measurement scales. We can then trade off a loss in one value dimension for a gain in another value dimension. The advantage of this approach is that it avoids Arrow's theorem by assuming unit commensurability, and it does so without the need of expressing all values in terms of money, which is an advantage compared to cost-benefit analysis.

It is worth noting that in the examples discussed in section "Value Conflict in Engineering Design," all relevant values are not (yet) measured on an interval scale. In the safety belt example, Table 1 represents measurements of the options on both the value of safety and the value of freedom on an ordinal rather than an interval scale. In this case, it might be possible to measure safety on an interval scale (by expressing it, e.g., in a measure of probability of death or injury); for the value of freedom, this seems much more difficult. When we look at the coolants example (example 3), in Table 3, environmental sustainability is operationalized in a measurement on two ratio scales (ODP and GWP), while health and safety are in the table measured on an ordinal scale. We can, however, also operationalize these latter values in such a way that they can be measured on interval scales (see Table 5).

To make value trade-offs, we do not only need an interval (or ratio) scale measurement of the individual values but also unit commensurability. To achieve

**Table 5** Properties of refrigerants. The data for OEL and LFL are based on ASHRAE (2013)

| | Environmental sustainability | | Health | Safety |
|---|---|---|---|---|
| | ODP | GWP | OEL (occupational exposure limit)[a] | LFL (lower flammability level)[b] |
| CFC 12 | 1 | 10,900 | 1,000 | None |
| HFC 134a | 0 | 1,430 | 1,000 | None |
| HFC 152a | 0 | 124 | 1,000 | 48,000 |
| HC 290 (propane) | 0 | 3 | 1,000 | 21,000 |
| HC 600a (isobutane) | 0 | 3 | 1,000 | 16,000 |

[a]OEL is "the time-weighted average (*TWA*) concentration for a normal 8-h workday and a 40-h workweek to which nearly all workers can be repeatedly exposed without adverse effect" (ASHRAE 2013, p. 4). It is measured in ppm (parts per million) v/v

[b]The minimum concentration in air at which flame propagation occurs. It is measured in ppm (parts per million) v/v

that, the decision-maker (designer) needs to be able to answer questions like "how many units decrease in GWP compensate for one-unit decrease in LFL?"[14] One problem in answering such questions is that trade-offs may not be constant over the entire domain. Consider, for example, the trade-off between costs and safety in the design of a car. It may well be that at low levels of safety, one is willing to pay more for a one-unit increase in safety than at higher levels of safety. So if one establishes value trade-offs, it should be done, taking into account the current value of values being traded off. Keeney (2002) discusses this and other pitfalls in making value trade-offs.

Apart from such avoidable pitfalls, the assumption of unit commensurability in making trade-offs raises the more fundamental issue that I also discussed in relation to cost-benefit analysis, namely, can a gain in one value dimension always compensate a loss in another dimension? As indicated, it has been suggested that unit incommensurability and a resistance to certain trade-offs are constitutive of certain values or goods like friendship. It has also been suggested that values may resist trade-offs because they are "protected" or "sacred" (Baron and Spranca 1997). Such trade-offs between protected or sacred values have also been called taboo trade-offs (Tetlock 2003).

Taboo trade-offs create an irreducible loss because a gain in one value cannot compensate or cancel out a loss in the other. The loss of a good friend cannot be compensated by having a better career or more money. One possible explanation for the existence of taboo trade-offs is that protected values correspond with moral obligations (Baron and Spranca 1997), i.e., they express an obligation to meet a certain value to a certain minimal extent. If interpreted thus moral obligations

[14]Note that GWP should be as low as possible, while LFL should be as high as possible.

define thresholds for moral values. It seems plausible that below the threshold, the moral value cannot be traded off against other values because the moral obligation is more or less absolute; above the threshold, trade-offs may be allowed.

## Maximin

What if we lower our assumptions about what information is available, i.e., if we do not longer assume the possibility of ratio scale measurement of value (as in cost-benefit analysis) or of unit commensurability (as in trade-offs)? If we still assume some form of commensurability, i.e., what we called level commensurability, a decision rule known as the maximin rule becomes possible. This decision rule tells us to select that alternative that scores best, compared to the other alternatives, on its lowest-scoring value. The advantage of this approach is that it avoids Arrow's theorem (by assuming level commensurability) without assuming unit commensurability and, therefore, without the need for trade-offs. This advantage comes, however, at a certain price as we will see.

Consider again the safety belt case. If we were to compare the traditional safety seat belt and the automatic seat belt with the maximin rule, we are to proceed as follows. First, we judge on what value each of the alternatives scores worst (compared to the other values on which we score that alternative). For the traditional safety seat belt, the worst-scoring criterion is most likely safety, and for the automatic seat belt, it is most likely freedom. In comparing the two alternatives, we should now answer the question: What is worse, the score of the traditional seat belt on safety or the score of the automatic seat belt on freedom? If we answer the latter, we should choose the traditional seat belt. (We can then repeat the procedure to compare the winning alternative with the seat belt with warning signal.)

As the example suggests, for the maximin rule, we only need ordinal measurement of the relevant values. In this respect, it is less demanding than the previous two approaches. At the same time, the judgments that this approach asks us to make seem quite complicated, as it asks us to compare the scores of two alternatives on *different* value dimensions; more formally, the method asks us to compare the score of option a on value v with the score of option b on value w. Especially if there are many alternatives, this may be hard and cumbersome, if not impossible.

One may also wonder how sensible the maximin rule is as a decision rule for conflicting values in engineering design. If we try to interpret what the rule means in the context of engineering design, it boils down to what may be called strengthening the weakest link. One selects the design in which the weakest link of that design (i.e., the worst-scoring value) is relatively the strongest compared to the alternatives. Such an approach seems especially sensible if one wants to avoid underperformance on any of the relevant values (or design criteria). We may therefore say that the maximin rule results in a kind of "robust design."

It should be noted, however, that in some situations, the maximin rule leads to seemingly irrational results. Suppose I have a seat belt design that scores worse on safety than on freedom. Now suppose that through some tinkering, I develop a

design that scores only very slightly worse on safety but much better on freedom than the original design. Obviously, this new design will also score less on safety than on freedom. The maximin rule now tells us to prefer the first design over the new design whatever small the loss in safety (as long as there is some nonzero loss) and whatever big the gain in freedom. At least in some occasions, this seems the wrong advice.

## Satisficing

All previous approaches aim at selecting the best alternative (although they define the best differently, especially in the case of maximin). In contrast, in satisficing, one does not look for the optimal option, but first sets an aspiration level with respect to the options that are good enough and then goes on to select any option that exceeds that aspiration level (Simon 1955, 1956). Designers are reported to be satisficers in the sense that they set threshold values for the different design requirements and accept any design exceeding those thresholds (Ball et al. 1994). So conceived, satisficing may also be seen as a way of dealing with conflicting values, i.e., by setting thresholds for each value and then selecting any option that exceeds those thresholds.

An example of satisficing is to be found in the earlier-discussed case of the design of new refrigerants (example 2). On the basis of Fig. 1, the engineers McLinden and Didion drew a more specific figure with respect to the properties of CFCs, which is shown as Fig. 2.

According to McLinden and Didion, the blank area in the triangle contains refrigerants that are acceptable in terms of health (toxicity), safety (flammability), and environmental effects (atmospheric lifetime). This value judgment is a type of satisficing because by drawing the blank area in the figure, McLinden and Didion – implicitly – establish threshold values for health, safety, and the environment.

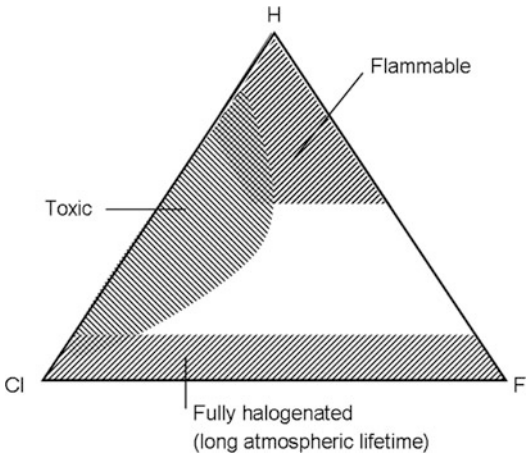



**Fig. 2** Properties of refrigerants (Figure from McLinden and Didion (1987))

**Table 6** Satisficing thresholds (implicitly) used by McLinden and Didion in drawing Fig. 2 and the score of the various options on these thresholds

| | Environmental sustainability | Health | Safety |
|---|---|---|---|
| Threshold | At least one H atom | Toxicity class A (not B) | Flammability class 1 (not 2 or 3) |
| CFC 12 | – | + | + |
| HFC 134a | + | + | + |
| HFC 152a | + | + | – |
| HC 290 (propane) | + | + | – |
| HC 600a (isobutane) | + | + | – |

Table 6 lists the thresholds that they set and it shows that of the earlier-considered alternatives only one, HFC134a, meets all thresholds.

Compared to the earlier-discussed methods, the main advantage of satisficing is that it requires less information, as it does not require any form of commensurability. The price to be paid is that it does not meet all of Arrow's requirements. In particular, it does not meet the condition of global rationality; rather, it orders the total set of alternatives into two sets, one with acceptable alternatives (i.e., those that meet all thresholds) and one with unacceptable alternatives (those that do not meet at least one threshold). Sometimes, the set of acceptable alternatives might consist of one item, as is the case in Table 6, and then it is clear what alternative to choose. However, the set of alternatives may also contain more than one alternative or be empty. If there is more than one acceptable alternative, the decision problem has not been solved yet. To be able to select one alternative, we might opt to satisfice with stricter thresholds or opt for one of the other approaches. If the set of acceptable alternatives is empty, we need to decide whether it is perhaps better not to design something (as no alternative meets all thresholds) or whether the thresholds should perhaps be relaxed.

As the above discussion already suggests, the core issue in satisficing is how to set thresholds. If thresholds are set in an arbitrary way, satisficing can hardly be seen as a rational method for dealing with value conflicts. However, in some situations, thresholds can be based on information external to the decision problem (cf. Van de Poel 2009). They may, for example, be based on technical codes and standards. This indeed happened in the refrigerants case discussed above: the thresholds for both toxicity and flammability were based on equivalence classes and thresholds that were defined in relevant codes and standards such as the ASHRAE Code for Mechanical Refrigeration.[15]

[15]For a more detailed discussion, see the chapter "► Design for Values and the Definition, Specification, and Operationalization of Values."

Thresholds may also be based on the law or company policy. One particular interesting possibility is to base thresholds on moral obligations. Earlier, I suggested that so-called taboo trade-offs may be due to the fact that some values can, for moral reasons, not be traded off below a certain threshold, as meeting a threshold corresponds with some moral obligation. In such cases, thresholds may thus be based on moral obligations (although it may be hard to define exactly where the threshold is between meeting and not meeting a moral obligation). So applied, the satisficing decision rule has the big advantage that it avoids the choice of a morally unacceptable alternative.

It should be noted that if thresholds are based on external information, it is likely that in many cases, satisficing will not lead to the selection of just one alternative. Especially if more than one alternative is still considered acceptable, between which external thresholds cannot decide, it seems most reasonable to use one of the other discussed approaches. In that sense, satisficing is an approach that is maybe best combined with other approaches.

Finally, just like the maximin rule, satisficing may sometimes result in (seemingly) irrational results. Suppose we have a refrigerant that meets the thresholds for safety (e.g., expressed in LFL) and environmental sustainability (e.g., expressed in GWP). Now suppose we find another refrigerant with a much lower GWP (so much better in terms of environmental sustainability) and a little worse in LFL (i.e., in terms of safety). Now also suppose that the decrease in safety is small but big enough to just fall below the threshold. Satisficing with the given thresholds now tells us that the second option should never be preferred to the first (as it does not meet all thresholds) whatever the gain in terms of environmental sustainability. Again, at least occasionally, this seems the wrong advice.

## Judgment: Conceptualization and (Re)specification

We will now look at an approach that emphasizes judgment and reasoning about values. This approach aims at conceptualizing and (re)specifying the values that underlie the conflicting design criteria. The advantage of this approach is that it might solve a value conflict while still doing justice to the conflicting values and without the need to make the values commensurable or to define thresholds for them.

The first thing to do when one wants to exercise judgment in cases of trade-offs is to identify what values are at stake in the trade-off and to provide a conceptualization of these.[16] What do these values imply and why are these values important? Take the value of freedom in the case of safety belts. Freedom can be conceptualized as the absence of any constraints on the driver; it then basically means that people should be able to do what they want. Freedom can, however, also be valued as a necessary precondition for making one's own considered choices; so conceived freedom carries with it a certain responsibility. In this respect, it may be argued that

[16]This paragraph draws from Van de Poel (2009).

a safety belt that reminds the driver that he has forgotten to use it does not actually impede the freedom of the driver but rather helps him to make responsible choices. It might perhaps even be argued that automatic safety belts can be consistent with this notion of freedom, provided that the driver has freely chosen to use such a system or endorses the legal obligation for such a system, which is not unlikely if freedom is not just the liberty to do what one wants but rather a precondition for autonomous responsible behavior. One may thus think of different conceptualizations of the values at stake, and these different conceptualizations may lead to different possible solutions to the value conflict. Some conceptualizations might not be tenable because they cannot justify why the value at stake is worthwhile. For example, it may be difficult to argue why freedom, conceived of as the absence of any constraint, is worthwhile. Most of us do not strive for a life without any constraints or commitments because such a life would probably not be very worthwhile. This is not to deny the value of freedom; it suggests that a conceptualization of freedom only in terms of the absence of constraints misses the point of just what is valuable about freedom. Conceptualizations might not only be untenable for such substantial reasons, they may also be inconsistent or incompatible with some of our other moral beliefs.

To make values operative in design, they do not only need to be conceptualized but also to be *specified*.[17] Specification is the translation of a general value or norm into more specific design requirements. The requirement can be more specific with respect to (a) scope of applicability of the norm, (b) the goals or aims strived for, and (c) actions or means to achieve these aims (cf. Richardson 1997, p. 73). An example is the specification of the value of safety into the following design requirement: "minimize the probability of fatal accidents (specification of the goal) when the chemical plant is operated appropriately (specification of the scope) by adding redundant safety valves (specification of the means)." In this case, the design requirement specifies the general norm in three dimensions, but specification may also be restricted to one or two dimensions.

A specification substantively qualifies the initial value or norm by adding information "describing what the action or end is or where, when, why, how, by what means, by whom, or to whom the action is to be done or the end is to be pursued" (Richardson 1997, p. 73). Obviously, different pieces of information may be added so that a general value or norm can be specified in a large multiplicity of ways. Not all specifications are adequate or tenable, however. In general, one would want to require that actions – or in our case, designs – that count as satisfying the specific design requirements also count as satisfying the general value or norm (cf. Richardson 1997, pp. 72–73). In the above example, "safety" is specified as "minimizing the probability of fatal accidents." This specification is adequate if in all cases in which the probability of fatal accidents is minimized, safety is maximized. Now arguably, safety encompasses not only avoiding or at least minimizing fatal accidents but also avoiding or minimizing accidents in which people get hurt

[17]This and the next paragraph draw from Van de Poel (2013).

but do not die. This does not make the specification necessarily inadequate, however. Maybe, it is known on the basis of statistical evidence, for example, that in this type of installation, there is a strict correlation between the probability of fatal accidents and the probability of accidents only leading to injuries, so that minimizing the one implies minimizing the other. In that case, the specification may still be adequate. In other situations, it may be inadequate and it might be necessary to add a design requirement related to minimizing nonfatal accidents.

Usually, more than one specification of a value will be tenable. This offers opportunities for dealing with value conflict in design. Value conflicts in design are in practice always conflicts between *specifications* of the values at stake because abstract values as such are too general and abstract to guide design or to choose between options. So if there is room for different possible specifications of the values at stake, it might be possible to choose that set of the specifications of the various values at stake that are not conflicting. Sometimes, it will only become apparent during the design process, when the different options have been developed and are compared that certain specifications of the values at stake are conflicting. In such cases, it may sometimes be possible to respecify the values at play so as to avoid the value conflict.

An interesting example of respecification took place in the refrigerant example 3.[18] In the first instance, the industry preferred HFC134a as alternative to CFC 12, basically following the satisficing reasoning as explained in the previous section (see also Table 6). However, environmental groups were against this alternative as they viewed the threshold for environmental sustainability (at least one H atom) too lenient, especially because it resulted in much higher GWPs (global warming potentials), than if a flammable coolant was chosen. At some point, Greenpeace succeeded in convincing a former East German refrigerator producer of using a flammable coolant in its new design. The refrigerator was also able to acquire the safety approval of the German certification institute TUV. Following the success of this refrigerator, German and later other European refrigerator producers also switched to flammable coolants like propane and isobutane. Such coolants were seen as acceptable despite their flammability because a new specification of safety was developed. Where safety was first specified as nonflammability of coolants, it now came to be specified as a low explosion risk of the whole refrigerator. It turned out to be possible to achieve a low explosion risk even with flammable coolants.

Although it might be possible to solve a value conflict in design through respecification, this will not always be possible. Even in cases in which it is possible, it may not always be desirable. It may especially not be desirable if respecification leads to a serious weakening of one of the values compared to the original specification (Hansson 1998). Still, solving a value conflict through respecification does not necessarily or always imply a weakening of one of the values (Van de Poel forthcoming).

[18]Based on van de Poel (2001). See also the chapter "► Design for Values and the Definition, Specification, and Operationalization of Values."

## Innovation

The previous approach treats the occurrence of value conflict merely as a philosophical problem to be solved by philosophical analysis and argument. However, in engineering design value conflicts may also be solved by technical means. That is to say, in engineering it might be possible to develop new, not yet existing, options that solve or at least ease the value conflict. In a sense, solving value conflicts by means of new technologies is what lies at the heart of engineering design and technological innovation. Engineering design is able to play this part because most values do not conflict as such, but only in the light of certain technical possibilities and engineering design may be able to change these possibilities. An interesting example is the design of the storm surge barrier in the Eastern Scheldt estuary in the Netherlands (example 2) that eased the value conflict between safety and ecology.

The reason why technical innovation can ease value conflicts is that it enlarges the feasibility set. This is a clear advantage of this approach. Often, however, technical innovation will not lead to *one* option that is clearly better than all others, so that choices between conflicting values still have to be made. In this respect, innovation only presents a partial solution to value conflicts in design.

According to van den Hoven et al., "technical innovation results in moral progress in those cases in which it means an improvement in *all* relevant value dimensions" (Van den Hoven et al. 2012, p. 152). Of course, not all technical innovations imply an improvement in all relevant value dimensions. Sometimes, a gain in one value dimension comes at the cost of a loss in another value dimension. Sometimes, the technical innovation creates new problems or side effects, which require new value dimensions to be taken into account. Sometimes, the technical innovation only addresses the initial problem in as far as it is amendable to a technological solution. It might also be that the values themselves change due to technical development; an often mentioned example is the change in sexual morality due to the development of anticonceptives. Technical innovation may also create new choices and dilemmas, as in the case of prenatal diagnosis, that we do not want to have.

Pointing at technical innovation as a way to deal with value conflicts does not yet make clear how to develop the kind of innovations that actually eases value conflicts. One approach here may be to translate the values into more specific design requirements that can guide design (Van de Poel 2013). Another interesting approach is that of value dams and value flows that have been proposed in VSD (see chapter "► Value Sensitive Design: Applications, Adaptations, and Critiques"). A value dam is a technical feature that is (strongly) opposed by one or more stakeholders because it conflicts with important values; a value flow is a technical feature that is for value reasons supported by various stakeholders. So, a value dam tells where not to go in innovation, while a value flow suggests technical features that should be included. In the case of the Eastern Scheldt, a design feature like "complete closure" can be associated with a value dam given the strong opposition from environmental groups, but also the design feature "no dam" met strong opposition from the government agency Rijkswaterstaat and therefore also

constituted a value dam. The design features "half-open" and "flexibly open/closed" on the other hand constituted value flows as they allowed meeting both the values of safety and ecology.

## Comparison of Methods and Conclusion

Above, I discussed six methods for dealing with value conflict in design. We saw that each method has its pros and cons; this has been summarized in Table 7. As the table shows, no method in isolation provides a complete solution to the problem of value conflict in design. It might, however, be possible to combine methods and so achieve an acceptable procedure for dealing with value conflict in design.

In particular, different methods may be required for value conflicts that amount to moral dilemmas than for value conflicts that do not entail a moral dilemma. As we have seen in section "Value Conflict in Engineering Design," not all value conflicts entail moral dilemmas but some do. Value conflicts amount to a moral dilemma if there are values at stake that correspond to moral obligations that cannot be simultaneously met. I have above suggested that such moral obligations may be characterized as a minimal threshold that should be met on each (or at least some) of the relevant value dimensions. Figure 3 represents this idea. For each of the values, a minimal threshold has to be met to meet moral obligations. We can now define the moral opportunity set as the set of options that is feasible and meets all minimal thresholds. If the moral opportunity set is empty, we are confronted with a moral dilemma.

It should be noted that satisficing might help to ensure the choice of an option within the moral opportunity set if that set is nonempty, although it cannot choose between options within the moral opportunity set. If the moral opportunity set is empty, innovation is a particularly attractive option because, as suggested by Van den Hoven et al. (2012), it may make the moral opportunity set nonempty.

On the basis of these ideas, I want to end with a particular suggestion of a stepwise approach that combines the methods for cases of conflicting *moral* values:

1. *Satisficing with moral obligations*. The goal of this step is to rule out morally unacceptable options. To do so, one looks for moral obligations that correspond to the relevant moral values and judges whether these correspond with (minimal) thresholds to be met by those values.

   This step requires *judgment* in order to identify the moral obligations and to define the corresponding thresholds. It is important to focus on moral obligations rather than on other external constraints that may (also) set threshold values because these other constraints may not be morally desirable. It is also advisable to focus on clear and uncontroversial moral obligations as this step is meant to rule out clearly morally unacceptable options.

   For setting the thresholds, it is also advisable to define tangible attributes for the values that can be assessed in design. This will require a specification of the relevant values.

**Table 7** Overview of methods for dealing with value conflicts in design

| Method | How are the values weighted? | Main advantages | Main disadvantages |
|---|---|---|---|
| **Cost-benefit analysis** | All values expressed in monetary terms | Values are made commensurable on a common scale so that the best option can be chosen | Requires ratio measurability and value commensurability to avoid Arrow's theorem |
| | | | Various fundamental and methodological problems with expressing values in money |
| **Direct trade-offs** | Trade-offs between the different values | "Best" option chosen ("optimal design") | Requires unit commensurability to avoid Arrow's theorem |
| | | | Making trade-offs may not always be (morally) acceptable |
| **Maximin** | Comparison between scores on various values | Selects a design that scores best on worst criterion ("robust design") | Requires level commensurability to avoid Arrow's theorem |
| | | | May sometimes lead to (seemingly) irrational choices |
| **Satisficing** | A threshold is set for each value | The selected alternatives meet the thresholds | How to set thresholds in a rational and acceptable way? |
| | | Can help avoid morally unacceptable alternatives | Subject to Arrow's theorem: criterion of collective rationality not met |
| | | No direct trade-off between the criteria | May sometimes lead to (seemingly) irrational choices |
| **Judgment** | Conceptualization and specification of the various values | Might solve value conflict by judgment and (re) specification | Not all value conflicts can be solved in this way |
| | | | Respecification may lead to an unacceptable weakening of moral obligations |
| **Innovation** | Not applicable | Can lead to alternatives that are clearly better than all of the present alternatives | Does not solve the choice problem in many cases |

After satisficing with the so-defined thresholds, the moral opportunity set may either be empty or not. If it is empty, innovation may be required to look for solutions that make the moral opportunity set nonempty. If the moral opportunity set is nonempty, innovation is still advisable because it might be possible to find better options than the currently available. In both cases, the next step therefore is:

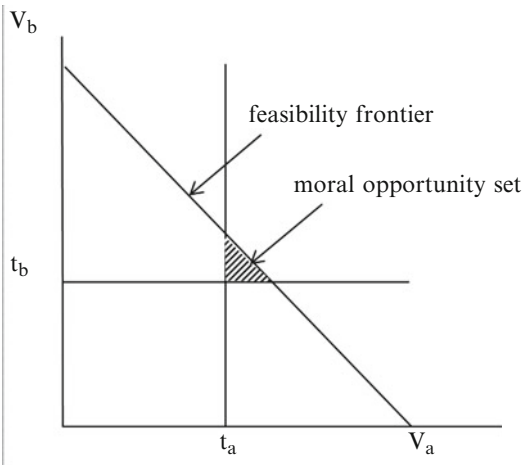



**Fig. 3** The moral opportunity set. $V_a$ and $V_b$ are the values at stake, and $t_a$ and $t_b$ the minimal thresholds corresponding to moral obligations for these values. The blank area left of the feasibility frontier is the feasibility set. Note that depending where the feasibility frontier is, the moral opportunity set may be empty, in which case, we are confronted with a moral dilemma (The figure is based on Van den Hoven et al. (2012))

2. *Innovation*. The goal of this step is to develop new options that better meet the relevant values. Doing so might require a further *specification* of the relevant values in order to be better able to develop new options that better meet those values. Also, other VSD tools like value dams and value flows might be helpful here.

   After this step, there are roughly three possibilities: (1) the moral opportunity set is (still) empty and one should choose whether to design a product or not, (2) the moral opportunity set is nonempty and contains exactly one option, and (3) the moral opportunity set is nonempty and contains more than one option. In the second case, no further step is required. In the first and third cases, the next step is a choice, but as the nature of the choice is somewhat different, I differentiate between two versions of the next step.

3a. *Choice* (if the moral opportunity set is empty). As there is no design option that meets all relevant moral obligations, one should wonder whether to design a product or not. Depending on the specific case, it might be possible to solve the moral dilemma through respecification, i.e., in such a way that there is a design that meets all minimal moral thresholds. However, one should take care not to unacceptably play down moral obligations or values in respecification.

3b. *Choice* (if the moral opportunity set contains more than one option). If the moral opportunity set contains more than one option, a choice has to be made between these options. The most appropriate approaches for doing so are cost-benefit analysis, direct trade-offs, and maximin, as the other three approaches do usually not narrow down the choice to one option. Of these three approaches, direct trade-offs often seem to me the most desirable. The reasons for this are as follows. In our case, the options among which a choice needs to be made all meet minimal thresholds set by moral obligations (as a consequence of step 1); this makes, as earlier suggested, trade-offs between values in most cases acceptable. Compared to cost-benefit analysis, direct trade-offs have the advantage of being less informationally demanding, and it lacks the

disadvantages that come with expressing values in monetary units. If we compare maximin with direct trade-offs, one might think that maximin is less informationally demanding; on the other hand, as we have seen, the level commensurability that is required for maximin requires quite complicated judgments. Moreover, maximin may occasionally lead to (seemingly) irrational choices.

I do not want to suggest that the above approach is the only way to combine the various methods for dealing with conflicting values in a reasonable way. There may be other ways of doing it. My proposal also specifically is meant for conflicts between *moral* values, and value conflicts between nonmoral values (or between moral and nonmoral values) may be better treated differently. Moreover, I have been assuming in this contribution that the designer takes a moral point of view. This assumption may not always be realistic, and even from a moral point of view, the designer may not always be required to do what is morally best, as it may be good enough to choose an option that is morally acceptable but perhaps not morally best.

## Cross-References

- ▶ Design for the Value of Safety
- ▶ Design for the Value of Sustainability
- ▶ Design for Values and the Definition, Specification, and Operationalization of Values
- ▶ Value Sensitive Design: Applications, Adaptations, and Critiques

## References

Arrow KJ (1950) A difficulty in the concept of social welfare. J Polit Econ 58:328–346

Arrow KJ, Raynaud H (1986) Social choice and multicriterion decision-making. MIT Press, Cambridge, MA

ASHRAE (2013) Designation and safety classification of refrigerants, ANSI/ASHRAE standard, 34-2013. ASHRAE, Atlanta

Ball LJ, Evans JSBT, Dennis I (1994) Cognitive processes in engineering design: a longitudinal study. Ergonomics 37(11):1753–1786

Baron J, Spranca M (1997) Protected values. Organ Behav Hum Decis Process 70(1):1–16

Boyce RR, Brown TC, McClelland GH, Peterson GL, Schulze WD (1992) An experimental examination of intrinsic values as a source of the WTA-WTP disparity. Am Econ Rev 82(5):1366–1373. doi:10.2307/2117484

Franssen M (2005) Arrow's theorem, multi-criteria decision problems and multi-attribute preferences in engineering design. Res Eng Des 16:42–56

Hansson SO (1998) Should we avoid moral dilemmas? J Value Inq 32(3):407–416. doi:10.1023/a:1004329011239

Hansson SO (2007) Philosophical problems in cost–benefit analysis. Econ Philos 23:163–183

Horowitz JK, McConnell KE (2002) A review of WTA/WTP studies. J Environ Econ Manag 44(3):426–447. doi:10.1006/jeem.2001.1215

Hylland A (1980) Aggregation procedure for cardinal preferences: a comment. Econometrica 48(2):539–542. doi:10.2307/1911117
Jacobs JF, Van de Poel I, Osseweijer P (2014) Clarifying the debate on selection methods for engineering: Arrow's impossibility theorem, design performances, and information basis. Res Eng Des 25(1):3–10
Keeney RL (2002) Common mistakes in making value trade-offs. Oper Res 50(6):935–945. doi:10.2307/3088614
Kroes P, Franssen M, Bucciarelli L (2009) Rationality in design. In: Meijers A (ed) Philosophy of technology and engineering sciences, vol 9, Handbook of the philosophy of science. Elsevier, Oxford, pp 565–600
Levi I (1986) Hard choices. Decision making under unresolved conflict. Cambridge University Press, Cambridge
Mas-Colell A, Sonnenschein H (1972) General possibility theorems for group decisions. Rev Econ Stud 39(2):185–192. doi:10.2307/2296870
May KO (1954) Intransitivity, utility, and the aggregation of preference patterns. Econometrica 22(1):1–13
McLinden MO, Didion DA (1987) Quest for alternatives. ASHRAE J 29(11):32–42
Raz J (1986) Value incommensurability: some preliminaries. Proc Aristot Soc 86:117–134
Richardson HS (1997) Practical reasoning about final ends. Cambridge University Press, Cambridge
Roberts KWS (1980) Interpersonal comparability and social choice theory. Rev Econ Stud 47(2):421–439
Sen AK (1970) Collective choice and social welfare. Oliver & Boyd, Edinburg/London
Simon HA (1955) A behavioral model of rational choice. Q J Econ 69:99–118
Simon HA (1956) Rational choice and the structure of the environment. Psychol Rev 63:129–138
Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) (2007) Climate change 2007: the physical science basis: contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge/New York
Tetlock PE (2003) Thinking the unthinkable: sacred values and taboo cognitions. Trends Cogn Sci 7(7):320–324
Tinbergen J (1959) Sociaal-economische aspecten van het Deltaplan. In: Deltacomissie (ed) Rapport Deltacommissie, Bijdrage VI, Onderzoekingen van belang voor het ontwerpen van dijken en dammen. Staatsdrukkerij- en Uitgeverijbedrijf, Den Haag, pp 61–74
Tsui K-Y, Weymark JA (1997) Social welfare orderings for ratio-scale measurable utilities. Econ Theory 10(2):241–256. doi:10.1007/s001990050156
van de Poel I (1998) Changing technologies. A comparative study of eight processes of transformation Of technological regimes. University of Twente, Enschede
van de Poel I (2001) Investigating ethical issues in engineering design. Sci Eng Ethics 7(3):429–446
Van de Poel I (2009) Values in engineering design. In: Meijers A (ed) Philosophy of technology and engineering sciences, vol 9, Handbook of the philosophy of science. Elsevier, Oxford, pp 973–1006
Van de Poel I (2013a) Translating values into design requirements. In: Mitchfelder D, McCarty N, Goldberg DE (eds) Philosophy and engineering: reflections on practice, principles and process. Springer, Dordrecht, pp 253–266
Van de Poel I (forthcoming) Dealing with moral dilemmas through design. In: Van den Hoven J, Miller S, Pogge T (eds) The design turn in applied ethics. Cambridge University Press, Cambridge
Van de Poel I, Royakkers L (2011) Ethics, technology and engineering. Wiley-Blackwell, Oxford
Van den Hoven J, Lokhorst G-J, Van de Poel I (2012) Engineering and the problem of moral overload. Sci Eng Ethics 18(1):143–155. doi:10.1007/s11948-011-9277-z
Williams B (1973) Problems of the self. Philosophical papers 1956–1972. Cambridge University Press, Cambridge

# Design for Values and Operator Roles in Sociotechnical Systems

Maarten Franssen

## Contents



**Abstract**

Engineering is increasingly of systems that are not only complex – in being multilayered – but also hybrid – in containing people as components. This chapter does not discuss the already well-researched safety concerns generated by this development with respect to operators, users, and bystanders but instead addresses values of relevance to the presence in such systems of operators as agents: values associated with what we can reasonably ask people to do and what we make people responsible for. From the perspective of design, systems containing people as components deviate in four ways from traditional systems consisting of all-hardware components. (1) Such systems are not designed as a

M. Franssen (✉)
Section of Philosophy, Fac. TPM, Delft University of Technology, Jaffalaan 5, Delft, The Netherlands
e-mail: m.p.m.franssen@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_8

whole but gradually and modularly. (2) They are typically delivered as incomplete because their operators have to be added by the client or owner during implementation; only operator *roles* are strictly speaking designed. (3) Persons performing these roles become components of the system only partially; unlike hardware components their behavior continues to be monitored and controlled from a perspective external to the system. (4) Operator roles are often explicitly conceived as being partially external, in that an operator's task is ultimately to ensure that the system continues to function properly come what may. These features lead to conflicts with the autonomy of operators as persons and the well-foundedness of assigning particular responsibilities to them. The difficulties described here should make us rethink whether traditional engineering approaches to system design are adequate for such hybrid systems.

**Keywords**
Sociotechnical system • System design • Operator • Responsibility • Autonomy

## Introduction

That failing technology puts people's lives at risk is true almost everywhere on earth. But technology can fail in many different ways, which differ in the consequences they have and what we can do about it. The particular aspect at issue in this chapter is that certain complex, systemic forms of technology rely on human operators and that failures of these systems, which typically but not necessarily involve errors on the part of the operators, not only put these operators' lives at risk as it does the lives of other people but also put their reputation at risk by making them minimally causally responsible for the potentially disastrous consequences of failure, and thereby pose a burden on them for the rest of their lives or, if they do not survive, on their family and relatives. Let me give a few examples.

In July 2002, two aircraft collided in midair over Überlingen in southern Germany after the crew of one of them received conflicting instructions – both to descend and to climb – from two different sources and chose the wrong one to follow. In June 2009 the crew of a French Airbus failed to take control of the aircraft when the autopilot disengaged due to a loss of airspeed data and steered the aircraft into a straight downward course lasting several minutes until it finally crashed into the Atlantic Ocean. In both cases the crews were not flying these aircrafts for their own sake but were executing a task: flying several hundred passengers who had all paid for the service of being transported through air and of whom none survived. The crews were operators in a complex system; they were part of the machinery that their employers were using to generate a profit by offering this transportation service, and they were destroyed together with other parts of that machinery. A failure to execute their task correctly puts these operators at risk just as much as it puts the customers whom they service at risk. The control room operators whose handling of the control rods during a test led to the destruction of reactor no. 4 in Chernobyl in 1986 also paid with their lives, still failing to

understand how their actions "entirely in accord with the rules" could have had this disastrous outcome while they were perishing in a Moscow hospital.

For operators the dire consequences can extend well beyond the immediate disastrous event. The Danish air traffic controller, employed by the Swiss company Skyguide, whose late interference was responsible for the issuing of two conflicting instructions that led to the 2002 midair collision, was later murdered by a Russian citizen who lost his wife and two children in the crash. A Yugoslavian air traffic controller whose similar oversight led to an earlier midair collision near Zagreb in September 1976 was convicted to a prison sentence of 7 years, to be released only in response to a worldwide petition by air traffic controllers after having served 2 years of his sentence. And an Italian air traffic controller who directed ground traffic at the time of a deadly collision of an airliner with a smaller aircraft during takeoff at the airport of Linate near Milan in October 2001 was afterwards convicted to a prison sentence of 8 years, a conviction that was upheld after two appeals, even though the official investigation report had not singled out his actions as in any particular and blameworthy way a major cause of the accident.

That the human components of such systems in technology are a special and frequent source of error, with often disastrous consequences, is hardly an original observation and has been argued repeatedly (Perrow 1984; Whittingham 2004). Humans, moreover, fail in other ways than hardware components. Humans are sensitive to the boredom generated by repetition. Rules may not be taken seriously if they interfere with operational procedures or if they too often "cry wolf."[1] Such issues are by now well recognized and well (though perhaps less well) understood, and methods for dealing with them have been and are being developed, making up the field of human factors engineering. Human factors engineering is aimed, however, at the general improvement of the reliability and safety of engineering systems and thus the protection from harm of operators, users, and bystanders alike. Much has already been achieved in this respect. Amalberti (2001) classifies civilian air traffic system and the European railroad systems, from which the majority of examples in this chapter are drawn, as "almost totally safe transportation systems." My aim in this chapter is not to contribute to this literature, nor is my focus a further increase in the safety and reliability of these systems; though of course it may contribute to that and I should certainly hope it does.[2] My aim is rather to discuss how the inclusion of people in technical systems generates value issues for, or related to, specifically these people and to suggest ways in which these issues can be identified and addressed, if probably not entirely resolved, from a design perspective.

[1] ACAS, for instance, the automated system for the avoidance of collisions between aircraft which played a role in the Überlingen crash, is notorious for generating false alarms, in particular when an aircraft is climbing or descending and thereby approaches another aircraft cruising at an altitude just below or above the altitude for which the climbing or descending airplane is heading ([Eurocontrol]; Pritchett et al. 2012a, b).

[2] Amalberti (2001), however, particularly discusses the limits to such further increase.

In the following three sections, first the concept of a sociotechnical system is discussed; then an analysis is presented of the ways in which such systems, consisting partly of technical, engineered devices and partly of human operators, differ from traditional engineering devices and systems, especially from the point of view of engineering design; and finally some consequences for the status of operators as valuing and evaluated persons are discussed. With a few exceptions, empirical support is drawn from air traffic cases, but the analysis and conclusions I hold to apply beyond this particular field to all systems with humans as components.

## Complex Technological Systems with Human Components

Complex people-containing entities which are conceived and implemented as instruments to serve some purpose will be referred to in this chapter as *sociotechnical systems*. Although this term was coined over half a century ago to indicate a more restricted concept, this is how the term is currently most often used.[3] The "socio-" emphasizes not only the mere presence of people as components of the system but also the fact that in order to initiate and coordinate the actions of these people, instructions, rules and regulations, and similar social "mechanisms" play a role. The term is therefore more specific than other terms used in the same field, such as engineering systems (de Weck et al. 2011), where the emphasis is on their engineering complexity and not particularly on the inclusion of people.

The people that are, by design, components of such sociotechnical systems I will refer to as *operators*, whatever the character of the activities they are supposed to perform. This includes, in the realm of air transportation, pilots and other aircraft crew, air traffic controllers, check-in attendants, luggage handlers and other airport employees, air ticket sale personnel, and so forth. Since sociotechnical systems are conceived and implemented as instruments to serve some purpose, every sociotechnical system presupposes some intentional or quasi-intentional "system owner," who uses the system as an instrument to achieve this purpose, and also an object for the transformation of which the system is used. Together the sociotechnical system, its user, and its object form an *instrumental system* of a particular type, a type ultimately determined by its owner–user, irrespective of what the sociotechnical system's operators think they are participating in.[4] The user of a complex sociotechnical system as an instrument is here referred to as the "system owner" of the instrumental system so created in the sense that this user, by using the

[3]See on this notion of a sociotechnical system Vermaas et al. (2011), Ch. 5, or Franssen and Kroes (2009). For earlier uses, see, e.g., Kelly (1978).

[4]In Franssen (2014), I develop a systematic account of the notion of instrumental system. My use of the term "sociotechnical system" there is slightly different from its use here: there I use it for a particular type of instrumental system, with an instrument of a certain complexity, but including a user and the object the transformation of which is what the user wishes to achieve, whereas here I use it in a more restricted sense to mean just this complex instrument.

instrument in a particular way, determines the kind of instrumental system that comes into being by the use. But this user need not own the instrument in any legal sense, though this may certainly be the case, even when the instrument is a nationwide sociotechnical system. What, precisely, is owned in such a case is, however, a difficult question; it excludes, in modern societies at least, all persons who nevertheless must count as components of the sociotechnical instrument.

Due to the user-defined "momentariness" of instrumental systems, any sociotechnical system can, just as much as any tangible object, be operative as instrument in a wide variety of instrumental systems – as basically acknowledged (but insufficiently emphasized) by systems engineering. An air transportation system can be an instrument in a profit-generating system for a private (typically incorporated) owner–user. And it can be an instrument in a flying system or transportation-through-air system for a private (typically individual) user. Note that the two instruments in these systems do not coincide. As for the profit-generating system, each customer to whom a flight is sold is a component of this system's instrument; without paying customers, the system would be "running idle," that is, would not generate a profit for its owner–user (who is, by being transformed into a condition of increased wealth, also the system's object). The owner–user of the profit-generating system is in its turn a component of the flying system's instrument; without the owner–user in place, there would be no live airline company to be used as an instrument; it would malfunction by not responding to the customer's inserted "coin" or button pushing, let alone fly the customer to their destination. Although the two systems overlap completely, in that they are instantiated by the same complex entity existing in the world, the boundaries separating the various roles run differently, depending on who is bringing about which change in the world using what.

The preceding two (closely related) examples are of instruments in commercial *service-providing* systems. Similarly we can also have commercial *product-delivering* systems with sociotechnical systems as their instrument. Again we are dealing ultimately with encompassing profit-generating systems, more in particular profit-generating-through-the-design-manufacture-and-sale-of-engineered-products systems. Within such a system, we can distinguish, as its instrument, a product-generating-and-delivering system. For example, if an aerospace company, say, Northrop, is invited to design a new jet fighter, then within Northrop a complex instrument will be organized for the design, development, manufacture, testing, delivery, and maintenance of this aircraft. These are the sorts of systems that the discipline of systems engineering has traditionally been concerned with: the jet fighter as a (entirely physical) system and the (partly social) system that has to be put in place in order for a system like a jet fighter to be brought into existence successfully.

The notion of a technical or engineering system, as used in systems engineering, generally refers to the complex wholes that figure as the instruments of my encompassing notion of an instrumental system. However, this could be either a system as delivered to a client (a jet fighter sold to government) or an ongoing service-providing system (a national air force) or the organization that delivers – designs and builds – such a jet fighter or sustains its operation. Sage and Armstrong,

Jr. (2000) characterize systems defined, developed, and deployed by systems engineers as being either products or services (p. 2) and distinguish service-oriented (e.g., an airport), product-oriented (e.g., automobile assembly plant), and process-oriented (e.g., a refinery) systems. Each of the three examples is a system that has many people among its components. Buede (2009) is less clear. On the one hand, he conceives of systems in a traditional sense of things built by engineers, taking for granted, e.g., that they can be painted green. On the other hand, the wider notion of a system that brings such engineering products into being (and which definitely cannot be painted green) is presupposed. All Buede's examples, however, presuppose product-delivering systems (e.g., the system delivering the F-22 jet fighter).

Additionally, there are noncommercial regulating and monitoring systems, conceived as large nation-spanning if not world-spanning infrastructural systems, e.g., the world civil air traffic regulation system, or the national or, say, the European electric-power-providing system. These are not finally instruments in some profit-generating system – at least they are not primarily conceived as such – but rather provided by states for the benefit of their citizens. They can, however, include subsystems which can be so characterized, for instance, any particular power plant participating in a national or supranational electric-power-providing system.

What I argue in this chapter is that the design of sociotechnical systems differs crucially from the designing of all-hardware devices or systems such as a jet fighter or the Global Positioning System, and that these differences give occasion to issues with respect to design for value that uniquely concern sociotechnical systems. I distinguish four major differences:

1. Sociotechnical systems are not designed, assembled, and tested from scratch and as a whole but designed and deployed modularly. As a result, they grow and develop almost organically, with the corresponding consequences, foremost being prone to emergent behavior. Although emergent behavior occurs also in designed all-hardware devices, the possibilities of controlling for it are much greater there.
2. Even though sociotechnical systems are not designed, tested, and implemented as a whole, they are designed and tested to considerable extent and are conceptualized and monitored from a design perspective. This is required by the inclusion of engineered devices as system components. The human components cannot be treated in the same way, however. The tasks to be performed by people – to monitor and if necessary achieve coordination between various technical components or between the manipulations by external people ("users," "customers") and internal technical components – are instead designed as slots to be filled once the system goes operational. Filling these slots with people is, typically, not the responsibility of the designer but the prerogative of the user. Sociotechnical systems, then, emerge from the design-and-manufacture stage incomplete. Their operators are not furnished with them by the product-delivering company, finely tuned to interact optimally with the hardware as any component of a hardware product is, but have to be added by the client, the prospective owner–user, as part of the system's implementation.

3. The people who fill the slots and thereby become components of the system do not coincide with their role, as hardware components do, but perform their operator roles as persons and (in the current state of organization of the world) as citizens. In other words, they become components of the system under design/deployment only partially. Operators reflect on the performance of their role; their actions continue to be monitored and controlled from an external, personal perspective.
4. Operator roles are as a rule not exclusively local but also global: though placed at certain nodes defined by output-to-required-input profiles, an implicit background task is typically to ensure that the system continues to function properly. The capacity and disposition of operators to reflect on task performance and system performance are therefore not only acknowledged but even presupposed. Operators, or rather the people performing operator roles, are supposed to perform a global monitoring role as well, which requires a perspective on the system from the "outside."

Together, these differences have major consequences both for the sort of values that should be taken into account in any form of design decision making concerning such systems – from conceiving of them all the way to implementing and maintaining them – and for how these values should be taken into account. In the next section I discuss them in more detail one by one and illustrate them using examples of system failures.

## Four Distinguishing Characteristics of Sociotechnical Systems

### Modularity

The first difference is that sociotechnical systems are hardly if ever designed and implemented as a whole and from scratch. They are too big for this and too encompassing to allow for the necessary isolation. Rather they are extended through design: they grow almost organically by having designed modules added to them or by having parts of the system replaced by designed modules. As a result, there is no single controlling instance monitoring the design of such systems, checking the progress toward meeting the design requirements, assessing the compatibility of the system components, and ultimately certifying the outcome. As a further consequence, coordination difficulties between system components and modules – the entities that are under unified design control – may crop up and must be expected. These difficulties are themselves not necessarily a consequence of the presence of operators in the system. An example that involved just hardware devices is the crash of the last Concorde on 25 July 2000 (see [Concorde], pp. 94–117). During takeoff, one of the tires of the landing gear blew out and a fragment of the tire bumped against one of the fuel tanks in the wings. The resulting shock wave in the full tank caused a hole of about 30 by 30 cm in the tank wall. The fuel escaping through the hole caught fire, and the scale and intensity of this fire

caused so much damage to the wing that the crew lost control of the aircraft. Now tire bursts had been a problem ever since the Concorde started flying, and damage to the wings as a result of these bursts had been a cause for concern.[5] Experiments were performed in 1980 to establish the amount of damage possible as a result of tire bursts, and it was concluded that no modifications of the wing tanks were called for. In 1982, however, it was independently decided that the wheels and tires of the landing gear needed strengthening, and as a result the Concorde's landing gear came to be equipped with thicker and heavier tires. This made the results of the 1980 experiments on likely damage to the wing fuel tanks due to tire bursts irrelevant, but the experiments were not repeated for the new, heavier tires. Accordingly an opportunity was lost to discover that more substantial damage caused by fragments of a burst tire had become a possibility, even though tire bursts themselves were now more seldom. This is a typical example where a device is modified after delivery by the company responsible for its design, when the arrangements for monitoring the design as up to standards are no longer in place.

Of course, this is just one source of failure, and the continued operation of the monitoring arrangement does not guarantee that all coordination issues between components are properly taken care of (as shown, e.g., by the notorious failure of the first Ariane 5 rocket in 1996 due to a failure to update a particular software module from a previous version of the rocket). The experimental study of the consequences of tire bursts undertaken for the Concorde shows an awareness of the importance of monitoring the coordination and interaction between all system components. However, once some of these components are people, that awareness seems often not to include them. A case that shows this is the midair collision mentioned in the opening section between a Tupolev 154 flown by Bashkirian Airlines (a local Russian airline) and a Boeing 757 cargo aircraft flown by DHL, which occurred over Überlingen in the south of Germany on 1 July 2002 (see [Überlingen], esp. pp. 69–71, 98–103). Normally air control should notice whether two aircraft are on a collision course and give instructions to one or both crews to change course. In this case, however, the Swiss air controller was distracted (by having to attend to two work stations at the same time, a situation aggravated by hardware problems and maintenance work going on) and failed to detect the potential conflict. For such cases, occasioned by previous midair collisions, an automatic and at the time of the accident obligatory airborne collision avoidance system (ACAS) was installed in both aircrafts. ACAS operates by having aircraft exchange signals and by generating automatic spoken instructions (or resolution advisories, abbreviated as RAs, in ACAS lingo) to the crews of the aircraft in a coordinated way, one crew receiving an instruction to descend and the other crew an instruction to climb. Due to the failure of the air controller to resolve the conflict

[5] Actually this damage was due to fragments of the aircraft dislodged by fragments of burst tires, never to fragments of tires directly.

in time, the ACAS of both aircrafts had been activated, first warning the crews of an approaching conflict and then generating an instruction to descend for the crew of the Boeing and an instruction to climb for the crew of the Tupolev. Just as this was happening, however, the air controller noticed his oversight and ordered the Tupolev to descend, ignoring the Boeing. The Russian crew received the contradictory instructions within just seconds of each other, the ACAS-generated instruction to climb coming in when the air controller had not yet finished telling the crew to descend. After some confusion, the Russian pilot decided to follow the air controller's instruction. Since both aircrafts descended, they remained on a collision course and eventually collided, resulting in the loss of 71 lives.

In this case, the necessity to coordinate interaction between two system components – air controller and ACAS – seems not to have been on the mind of anyone engaged with the system from a design perspective. ACAS has been developed in response to earlier midair collisions that were caused by air controllers failing to identify and resolve conflicts between aircrafts. It seems to have been designed and added from the perspective that where a human operator fails to "function correctly,"[6] an engineered device should be available as a remedy. It was apparently not considered, at least not by those people responsible for adding ACAS to the air traffic control system, that as long as the human operator remains there, its possible interference with the system through its interactions with other system components should be dealt with.[7]

Being unfinished, modular and extendible, due to which there are limits to the extent to which design is controlled from a single perspective, is not an exclusive feature of sociotechnical systems, as the Concorde case shows. Any technology that is important and expensive enough to merit continued monitoring with a view to redesign may have this character and be prone to the associated vulnerability. Neither is it an inevitable feature of sociotechnical systems. We can imagine a sociotechnical system to be designed and implemented from scratch as a whole, for instance in the case of a new country that, say, has emerged from a war, with the old infrastructure completely destroyed, and now providing itself with a power-providing system from scratch. Even then, however, due to the other major differences at issue here, sociotechnical systems are especially vulnerable to modular development because, in contrast to device systems like the Concorde, partial redesigns are possible at no immediate costs, as will be discussed below.

[6]What exactly constitutes a failing or malfunctioning operator is an interesting and important question, which I will not take up in this text, however. Somewhat more will be said on the topic in the next section.

[7]Neither the ICAO flight instructions for ACAS in force at the time of the accident nor the ACAS manufacturer's Pilots Guide (all cited in [Überlingen], pp. 51–53) considered the possibility of interference between ATC instructions and the automated ACAS process. See for a discussion of the Überlingen accident particularly from the viewpoint of the social technical hybridity of the system involved also (Ladkin 2004) and (Weyer 2006).

## Openness

The second major difference is that the human components of sociotechnical systems are not contained in them in the same way as hardware components are, manufactured to specifications, installed, tested, and fine-tuned, so as to function properly at delivery. Instead such systems are designed with "slots" to be filled by people during the deployment of the system. The system becomes operational only once all slots are filled. Actual coordination between all components will therefore become possible only in the deployment phase and will, as a consequence, hardly ever be achieved definitively.

The design of "slots" or "roles" for sociotechnical systems to some extent follows the design of all complex systems: the modularity of design, given the complexity of most devices and the fact that a wide spectrum of scientific disciplines is involved in designing particular components, brings with it that any designed system has the character of a number of slots linked by interfaces; any particular component can be characterized by a particular geometry and a particular input–output behavior such that it fits into the corresponding slot and, once there, will contribute through its behavior to the operation of the entire system. Similarly, any slot to be filled by a human operator will be bounded by interfaces to which a human person can connect by way of its senses and the parts under voluntary muscular control – typically the hands and feet. The required input–output behavior will be specified by a list of rules or instructions that, given the interfaces and the set of human capabilities presumed, serve to define the operator role in question.[8]

However, if this were all, we could strive to close the gap between purely technical systems and sociotechnical systems. Although people cannot be manufactured to specifications, people can be "worked upon" to exhibit the required input–output characteristics as closely as possibly as a lawlike pattern by *training* or by the more extreme form called *conditioning*. This approach to sociotechnical systems has been the ideal for one particular form of sociotechnology, the military, ever since the introduction of training and discipline as the "operational principle" in the Dutch army fighting the Spanish in the late sixteenth century by Prince Maurice of Nassau, whose extremely influential military innovations were conceived in close contact with engineers, including Simon Stevin.[9] The "scientific management" movement of the early twentieth century, often referred to as Taylorism, also approaches the human operator as a component like any other, whose behavior can be adjusted to fine-tune coordination between components and optimize the operation of the entire system.

[8]Note that the human capabilities presumed will typically be those of a normal adult person but need not be. Many occurrences of child labor were and are dependent on the child operators being smaller and more versatile than adults. But more extreme cases can be thought of. To give just one example, in the film *The Prestige* by Christopher Nolan, an elaborate stage illusion figures that requires the participation of several operators who must be blind for the trick to succeed.

[9]See in particular (Feld 1975) and (Kleinschmidt 1999), who both discuss the connection to technical systems.

Note that this is not, and does not amount to, slavery; rather it goes much further. Slavery is based on, or presupposes, a rational decision on the part of the slave to perform his or her "duties" given the consequences – punishment and death – if this is declined and often enough just as well given the rewards that adequate performance will bring. In this decision the slave exerts a form of autonomy. Slavery relies, therefore, on the recognition that generally persons do not coincide with the functional roles they perform in any system in which they enter – be it a social one or a sociotechnical one – and that performance of a role, the decision to do it, and how to do it are made from the underlying "platform" of the intentional, broadly rational person, as will be discussed further below. The effect aimed at by rigorous conditioning, in contrast, is ultimately to depersonalize the person who is executing a task and to separate the execution of the role from any monitoring by an "underlying" person. The commands in military drilling may look like instructions, to be understood as having some content and to require interpretation to disclose that content, but ultimately and ideally they are supposed to function as signals to which a corresponding conditioned response is expected. For military roles to opt for conditioning is understandable: in circumstances of war, no intentional, broadly rational person would decide to execute the role that a soldier is supposed to perform, and training and conditioning are not the only ways in which the owner–users of sociotechnical "war machines" such as armies have tried to overcome this problem, although in the sixteenth century they were very innovative ways.[10]

Due to the separation in sociotechnical systems between the system with its operator roles designed as slots to be filled and the people filling these slots, the system cannot be securely tested prior to deployment. Even if to some extent this is also true for hardware-only systems – where batteries, lamps, valves, fuses, tires, and anything else that is subject to wear and tear have to be replaced regularly – such components are sufficient constant, e.g., behave sufficiently lawlike, for this not to produce problems.[11] However, as a matter of fact, no amount of training and conditioning will prove sufficient to reach the amount of depersonalization that is required to make humans into sufficiently reliable, if still not completely reliable, deliverers of lawlike behavior. And even if it were sufficient, by far not a sufficient number of people could be found to volunteer for being converted into machine components in this way. And even if they could, legislation would block such a voluntary execution in many cases, just as it blocks in most countries the voluntary

[10]Lewis Mumford identified such instruments made of human beings more widely and termed them megamachines. See Mumford (1934). As for the introduction of military drill, Kleinschmidt (1999, pp. 609–611) especially stresses the effort toward a conditioned response to orders, in contrast to an interpretational response.

[11]Still things can go badly wrong due to underspecification of insertable components. An example is the explosion on the Plesetsk launch pad of a 8A92M Vostok space rocket in March 1980. The cause of the explosion was the use of soldering containing lead next to tin instead of pure tin soldering. The lead acted as a catalyst for the decomposition of the hydrogen peroxide component of the rocket fuel. See Varfolomeiev (2007).

sale of organs like the kidneys and blood; legislation often imposes on citizens the honoring of certain values – corporeal integrity, for one – against their autonomy.

A consequence of this feature of sociotechnical systems is that responsibility for and control of the exact operation and functionality of the system is shifted partly toward the owner–user of a sociotechnical system. That agent, by filling the operator slots with the persons who are going to perform the operator roles, decides how operators are trained and kept in shape and what the working conditions for each operator will look like, both generally and momentarily, and even has the final say on what the instructions defining each role will be. What this amounts to is that sociotechnical systems generally lack precise boundaries.[12]

The Chernobyl disaster furnishes a good – though perhaps extreme – example of what this may look like in practice. A first investigation into the causes, undertaken by a Soviet Union committee, pointed to the operators on shift during the test run which ended so dramatically as the persons to blame for it, due to their massive violation of basic procedures, a violation so massive that the probability of it occurring could not have been foreseen. What the committee judged to be violated, however, were procedures that they assumed were the procedures for handling the reactor, since these were the proper procedures on the basis of their expertise, knowledge of the reactor and of what the people responsible for its design had told them should have been the procedures. Later, however, the initial report had to be retracted when further investigation revealed that the operators had not violated a single procedure as they had in fact been laid down, in writing, by the global and local management of the responsible utility institution ([INSAG-7], pp. 1–2, 13–15).

A second example highlighting the importance of proper instruction and training as part of the functioning of a sociotechnical system is the loss of American Airlines Flight 587 in November 2001 ([AA587], pp. 133–156). The aircraft, an Airbus A300-600, crashed after repeated aggressive handling of the rudder in order to stabilize the aircraft in conditions of turbulence caused the entire rudder to separate. Violently working the rudder for this purpose was practice among American Airlines pilots, but although other aircraft could cope, the Airbus A300, which had a very sensitive rudder-operating mechanism, could not. Airbus industries had warned American Airlines not to use the rudder in this way in 1997, but pilot training courses were not affected. Accordingly, American Airlines was held liable for the accident.

As a consequence of the inclusion of operators, then, sociotechnical systems are, one could almost say, in permanent state of repair. Operator roles are filled by new

[12]Even if designer and owner–user are formally identical – say in the case of a state operating an infrastructure or a public utility – then as a user the state may have other interests and be under the pressure of other forces than when as designer. The notion of instrumental system, mentioned in the previous section, was particularly developed to clarify issues like these. As a mere instrument, a sociotechnical system is typically incomplete and "open." Only a full instrumental system, complete with its user and object-under-transformation, allows for the delineation of sharp boundaries of the full system and its major components.

people all the time, and even while in the system, people's abilities are not sufficiently constant to be left at the job unattended, so to speak. For this reason, no sociotechnical system can be expected to function properly without training for its personnel. Accordingly, not only should training procedures be included in role instructions but modules dedicated to training should be included in sociotechnical systems by design. Major failures often reveal serious shortcomings in this respect. The investigation following the head-on collision of two passenger trains at Ladbroke Grove, London, in October 1999, resulting in 31 casualties, concluded that the training of drivers as well as the training of signalers was defective. Especially there was no monitoring of training procedures at higher levels of the system, allowing trainers to proceed independently and as they saw fit, without adequate input on what to train for and how (see [Ladbroke Grove], §5.25-5.48 & §6.27-6.42). The consequences of this aspect for system performance and system design are more far-reaching, however. Amalberti (2001) argues that it explains why the performance of even "almost totally safe systems" like civilian air transport or rail transport cannot be improved beyond a certain limit: the modules dedicated to training need occasional failures as input in order to know what to train for. Since the circumstances of system operation change all the time, as do the capabilities and incapabilities of the people who act as its operators, the experience of how to train and what to train for is not asymptotic, and perhaps not even accumulative.

## Presupposition of Intentional Rule Following

If the two differences discussed until now were the only ones, sociotechnical systems could still be looked upon as approximating traditional engineering systems, where the human sciences are required only to deliver their best knowledge of effective methods of training and conditioning people but are not further required for describing and understanding the behavior of the system under design. Human operators could be treated as deliverers of input–output patterns similar to hardware components – although their reliability remained an issue, emphasizing that such systems would perhaps have to be seen as being in the prototyping stage indefinitely. There is something of this attitude in classical approaches to systems engineering, sometimes referred to as "hard systems thinking."[13]

However, and this is the third major form in which sociotechnical systems differ from traditional hardware systems, in the overwhelming majority of cases, people perform their roles as operators consciously, that is, they are conscious of the fact that they perform a role, instead of coinciding with that role, as they would when

[13]For the terminology, see (Checkland 1981). To be sure, animals trained for a conditioned response have been used as "hard" system components: e.g., B. F. Skinner's use of pigeons as component of the tracking mechanism of a missile guidance system.

being conditioned into performing the actions required by the role. Accordingly, performance of an operator role is a two-level process: operators must *understand* what is expected of them, depending on the circumstances, and they must *decide* to carry it out. In a situation of conditioning, there is no room for distinguishing between these two levels.

An operator's understanding of what is expected can be equated with the drawing up of an exhaustive list of instructions which can be seen as defining the role, although in practice skills and know-how that are acquired in education and training are also important, an aspect that is not further discussed here. To draw up these instructions and to implement a supportive training program is, then, as far as system design can get. The designers of sociotechnical systems generally have little or no control over the execution of these instructions. To secure adequate role performance, execution must be made worthwhile and rewarding, but whether it actually is depends on circumstances which lie to a large extent beyond the scope of system designers and are difficult to foresee. Operators reflect as persons on their role instructions, and they will judge the wisdom of doing what the role definition requires of them from their perspective of a person, who will generally perceive him- or herself to have clear interests. These interests may be judged to be harmed by some of the actions the person is expected to perform as an operator. There may also be vaguer goals and considerations the achievement or satisfaction of which may be jeopardized by such actions. How difficult it is to generalize here is shown by the Japanese kamikaze pilots during the Second World War but also by the behavior of the various operators present in the Chernobyl nuclear plant at the time of the explosion, showing both extreme self-sacrifice and extreme carelessness.

But just as a person performing as an operator does not coincide with the operator role but always performs the role as a person, just as little is a person restricted to the particular role he or she performs within the sociotechnical system at issue. It is in the nature of roles that one person can play or perform several roles at the same time and accordingly be committed to act on several different sets of instructions or rules at the same time. In fact, it is the standard situation in modern societies that people in their actions perform several roles at the same time, although some may be more in the background while one particular role is up front. For once, every adult acting as an operator is also a *citizen*, that is, is being held to act in accordance with the laws and regulations of the state under whose jurisdiction that person's actions fall.[14]

This creates several further challenges for the design for values of sociotechnical systems. For a start, system design must allow operators to be "good citizens" by

[14]It could be argued that citizenship is not a role because society is not an instrument in any instrumental system, at least not from the point of view of liberal democratic society. Whether or not this is granted, roles are anyway not confined to instrumental systems. Let citizenship then be a role in a social system, where I leave the precise meaning of "social system" intuitive, rather than an instrumental system; this does not change the situation.

not requiring them to act in violation of the legislation that they are subject to.[15] For a conflict, compare the regulation surrounding ACAS. The Überlingen midair collision was not the first incident of its kind. Almost a year earlier, over Tokyo, air traffic control also was late in noticing that two aircrafts were on collision course and interfered just when the ACAS system on board of both aircrafts has just generated its advisories to the crews (see [JA907/958], pp. 99–117). Here as well this resulted in one of the two crews receiving conflicting instructions. The crew of a Boeing 747 had received an instruction to descend by air traffic control, to which it responded immediately, only to receive 2 s later an ACAS instruction to climb. In response to the Überlingen collision, ICAO emphasized in the first update of its system of regulations that an ACAS instruction, once received, must always be followed, even when an explicit counter instruction by air traffic control is received. In this case, however, the pilot judged following the ACAS advisory to climb an extremely dangerous maneuver, given that he had already started to descend and was making a turn at the same time.[16] Acting according to the rule that an ACAS advisory must always be acted upon promptly, therefore, in this case conflicted with the rule that a pilot should operate an aircraft so as to secure the safety of all passengers.

To be sure, most regulations anticipate the possibility of such exceptional circumstances by adding an exception clause. In this particular case of ACAS, the ICAO document of flight procedures 8168 says "In the event of an RA pilots shall [. . .] respond immediately by following the RA as indicated, unless doing so would jeopardize the safety of the aeroplane" ([ICAO8168], p. III-3-3-1). Such an exception clause, however, is lacking in ICAO document 9863 dedicated entirely to ACAS. There, it is stated: "If an RA manoeuvre is inconsistent with the current ATC clearance, pilots shall follow the RA." Only three very specific exceptions to following an ACAS RA are mentioned: "Stall warning, wind shear and Ground Proximity Warning System (GPWS) alerts take precedence over ACAS RAs." ([ICAO9863], §5.2.1.14 & §3.21.1.10.) Not only may regulation occasionally betray a lack of anticipation of potential conflicts; it may even blatantly provoke it. The current British Rules of the Air declare themselves to apply "(a) to all aircraft within the United Kingdom; [. . .] (c) to all aircraft registered in the United Kingdom, wherever they may be." ([Rules of the Air 2007], p. 4.) These regulations inevitably impose a conflict upon any crew flying an aircraft registered in the United Kingdom in the airspace of other countries, where other rules of the air apply which may differ significantly from the British rules – and such differences exist with respect to such basic features as priority rules.[17]

[15]It is assumed here that these external rules, in particular national legislation, are morally in order.

[16]To add to the difficulty of the situation, the air traffic controller involved had, in his haste, made the error to select for an instruction to adjust its altitude the aircraft that was just making a turn rather than, as he should have done, the one that was flying a straight course.

[17]Quoted is the 2007 version still in force at the time of writing but under review for harmonization with the EU legislation, which is less "imperialistic."

These issues point to two major difficulties for the "closure under rules" of sociotechnical systems, which are both aspects of a crucial feature of sociotechnical systems already referred to above: the absence of sharp boundaries. The first is that system designers not only have limited control over operator instructions, as already mentioned, but additionally lack control over the content and the stability of any external set of rules with which operator instruction should be consistent. Not only are the rules that operators are legally obliged to follow drawn up by institutions that operate independently of system designers,[18] but these roles come into being through mechanisms that differ greatly from the practice of engineering. In contrast to traditional engineering systems, rules systems have no "developmental inertia": rules can change radically overnight, and these rule changes seem to be costless as well. Prima facie, at least, because there are long-term costs related to such changes, in the form of efficiency losses and damage to health and property due to inconsistencies in the rules, the effects of which emerge only in the course of time. This makes the design of sociotechnical systems a precarious affair. Although here as well there is some continuity with safety and health regulations that traditional hardware devices must satisfy: with respect to these, designers may face the difficulty that these regulations may change while you are designing for them. In such cases, however, whether or not a product does satisfy the rules is typically a clear yes/no matter. And once acquired a product may often still be used, even though producing it is no longer legal.

If we look only at the syntactic level, we may want the instructions defining an operator role to be appropriate, that is, leading to the optimal continued functioning of the system, as well as clear and unambiguous. Difficulties with both are to be expected, due to, again, the peculiar position of the role-defining rules in the total system design. Hardware components can be replaced by alternatives only with difficulty. The replacing itself is time consuming and expensive and may result in a breakdown. Replacement is also a specific interfering act with liability issues attached. Rules, on the other hand, can be changed at will: replacement is effortless and free and will hardly ever lead to an immediate breakdown of the system. Likewise liability works differently as far as regulation is concerned: in a delivery contract, it may often not be clear who is responsible for what rules and even when designer responsibility explicitly extends to rules the client may and typically will request some maneuvering space with respect to them.[19]

[18]They are, in a democracy at least, not totally independent, because systems designers are themselves citizens with voting rights and in this way and in other ways as well have the possibility of influencing political decision-making processes.

[19]Responsibility for the content of rules must be sharply distinguished from responsibility for the (non-)violation of rules once in force. As already discussed, the case of Chernobyl clearly showed the difference.

Certainly sovereign states will hold their power to legislate to have absolute priority.[20] Sociotechnical systems, however, may be so vast, and their owner–users so powerful that this priority claim can successfully be challenged. Additionally, they can be so vast that consistency with the regulation and legislation of many different countries is required. And there is no mechanism to secure the mutual consistency of these various national rule sets. Coordination mechanisms are usually not stronger than individual states; the absurdly imperial "Rules of the Air" of the United Kingdom quoted above are a case in point. The ICAO, in particular, being a UN agency, cannot impose regulation upon its member states. The most that its member states are obliged to do is to indicate whether or not they adopt ICAO regulation (see, e.g., [ICAOAnn2], p. 2–1).

## Combination of Internal and External Perspective

The second of the two difficulties mentioned above brings us to the fourth and final major distinguishing characteristic of sociotechnical systems. The exception clauses in operator instructions, far from being mere fillers for the holes left by the impossibility of drawing up precise instructions for every contingency, are actually a cornerstone of the design of sociotechnical systems containing human operators. Operator instructions are deliberately open-ended and leave room for interpretation because system users want to have their cake and eat it: operators should act in accordance with the rules drawn up by system design aimed to serve system performance best, but if necessary, when the system designers had it wrong for the particular circumstances or when unforeseen circumstances occur, they should adapt their handling of the system and thereby "save" continued system performance.

Designing for solutions where part of the control over the end result is held back from the designing engineers is a significant deviation from how engineers are used to conceive of and treat their "material." However, system design acknowledges it and builds on it. This is the second difficulty mentioned above, and it brings us to the fourth and final form in which sociotechnical systems differ from standard technical systems. If operators – more precisely the persons who perform operator roles – inevitably perform their roles from an ever-present background of being persons, which leads operators to reflect on the performance of the role and to judge the task's actions against the person's interests and all other rules that are perceived to have a say, then we may as well make use of that fact and make it work for the system.

[20]Many of these difficulties could then be thought to disappear if a state has the sole responsibility for design, implementation, and operation of a (e.g., infrastructural) system. As shown by the case of the Soviet Union, this situation is certainly not sufficient for solving the associated problems. Note that of all historical cases, the Soviet Union came closest to a state run by engineers; see, e.g., (Alexijewitsch 2013), p. 443.

Operators, then, are supposed to monitor the overall operation of the system and act to safeguard it, irrespective of whether the rules defining the specific operator role contain the corresponding instructions. The operator is even expected to act in violation of these instructions if this is necessary for securing the correct or adequate operation of the system. Operators are partly there to make up for "holes" in the system. Mitchell and Roberts describe the situation in the following terms (2009, p. 852):

> It is surprising how often system designers consider operators and their roles at the very *end* of the design process. Hardware is purchased, software is specified, and only then are the roles of operators considered – and that role is often to fill the gaps between computer-based subsystems. Designers normally think about the operator's role at a very high level: ensure everything works! [. . .] Typically the limits of technology define operator activities. In many contemporary systems, the prevalent design philosophy dictates that everything that *can be* automated *should be* automated. Operators are responsible for performing those activities that remain and, of course, ensuring that *all* activities, human *and* computer, are carried out effectively in the face of changing or unanticipated system and environmental conditions.

To some extent, legislation forces this perspective upon system designers and system owners. Legislation makes, for instance, the driver of a car or the pilot of an aircraft responsible for the safety of passengers and bystanders, irrespective of whether the driver or pilot is operating it privately or in a role as operator in a sociotechnical system and irrespective of the size of that sociotechnical system. What is more, legislation seems to share the gap-filling outlook on operators' responsibility to make up for the deficiencies of a system. A case in point is the prohibition in most countries of the operation of unmanned vehicles on public roads: a "driver" must be present who can take over in a case of emergency.

Together these four distinguishing features of sociotechnical systems set the stage for a consideration of value issues from a design perspective that concern the position of operators in such systems. How the system behaves, whether it behaves as designed and whether it can behave so, and in relation to that what its operators are supposed to contribute and how they are supposed to do so, affects their careers and their lives. The preceding discussion has argued that, in performing their roles, operators have to satisfy different and often conflicting requirements. From the top-down engineering perspective, people in operator roles are among the components of complex system, for the purpose of achieving and maintaining coordination or as (part of) the system's interface with users, who must "fit in" and behave in a specific way for the system to work. From the bottom-up societal perspective people are citizens who have a social and more generally moral responsibility for the results of their actions and for the things that they causally bring about, a moral responsibility which they typically have regardless of their position as operator in a system. And society may even formulate further responsibilities for people in particular operator positions on top of the general ones. As our society continues to rely on sociotechnical systems containing human operators and as engineering continues to be instrumental in sustaining this reliance, it is important to morally reflect on how the interests of these operators are cared for. This aspect tends to be

overshadowed by the pursuit of caring for the interests of the "general public," the people who are being served through these sociotechnical systems. In the next section, I discuss two aspects of this problem area: the retaining of a level of autonomy for operators and the responsible assignment of responsibility to operators.

## Design for Operator Values

Although engineering has spent much effort on designing systems such that the effects of human error can be contained and system performance is robust under human error, human error is still considered to be the major cause of failures. The circumstances that will generate a human error (from the standpoint of system performance) and the possible types of error are too various. The response of engineering has been and continues to be automation: the replacement of people by hardware components or, and increasingly, hardware-plus-software components. There is no question that automation generally increases system reliability and leads to safer system operation. Nevertheless many systems still contain human operators as components. The reliable automation of many tasks is still beyond current technical possibilities. Additionally, however, automation conflicts with the wish to impose on at least some operators an external perspective next to an internal perspective, as explained in the subsection "Combination of Internal and External Perspective." Operators are not only supposed to play their part in order to contribute to the system's functioning in the explicit way anticipated by the designers of the system but are additionally supposed to monitor this contribution and the system's ability to deal with circumstances the designers did not anticipate. These two aspects are not independent, of course: no need is felt for the continued presence of this external perspective exactly when designers are confident an automated, technical-only solution will do. And even if engineers could part with the urge to rely on human intelligence as a backup option, there is still the confidence of the public of (potential) customers in fully automated service systems to deal with. Plus the fact, already stated above, that legislation often imposes it.

Apart from these considerations of fail safety, there are also limitations to automation of a quite other type: certain sociotechnical systems are, by design, so open that they accept human-operated subsystems as ad hoc components. This is how the public road infrastructure of any country works. From the standpoint of any individual driver using it, the other drivers are human components of the system used, whose actions prepare the system in a precise, though constantly shifting, configuration for use. The price for this honoring of individual freedom is that the likelihood of (local) failure is much greater (cf. (Amalberti 2001), p. 111).

The ambiguous position of the operator is the main point of tension in the treatment of sociotechnical systems. This tension can be seen as a continuation of the dual perspective we have of humans: they are organisms, and as such falling under the descriptive vocabulary of science, but also persons, falling under the intentional and partly normative vocabulary of daily life. Included in the latter

perspective is our organized societal existence. Though the results of describing people as complex physical systems and researching their behavior in varying circumstances by scientific methods, plus the fruits of taking these results into account for design purposes, are undeniable; people do not appreciate being conceived of and treated as "mere" physical objects, as cogs and wheels. People value being seen and treated as persons who choose their actions on the basis of reasons. They prefer to understand, and to be able to defend, why the thing they are supposed to do is the correct thing to do and how it contributes to well-being. But there is another side to this coin: with intentionality and understanding comes the option of being held responsible and accountable by the imposed rules of social life when the contribution to well-being misfires. A key question, not addressed by human factors engineering, is how the design and implementation of systems which have both devices and humans as components should cope with this tension and what the possibilities are of relieving it. This section addresses two aspects of this question in the light of the preceding analysis of sociotechnical systems.

## Autonomy Versus Automation

To accept to become a component who is supposed to act according to a list of instructions is to deliver oneself to a situation where one is dependent on the quality of the information and the adequacy of the prescribed actions, without having a say in that quality and adequacy. In a more limited context – people working with electronic databases – Jeroen van den Hoven (1998) has characterized this situation as one of "epistemic enslavement." Not unlike a slave, an operator is "at the mercy" of the system he or she is a component of. The operator trades in his or her personal autonomy and rational and moral control of action, for, ultimately, that of whoever can count as the user of the system for some purpose.[21]

In the present context, where we look at systems regardless of whether the components are people or things, it should be noted that the condition of epistemic enslavement is not so new or exceptional as it may seem at first: it is a characteristic of any hierarchical social system. When acting "under orders" or "upon request," a person generally acts without knowing the reasons on the basis of which the action is required or justified. In a sense, then, such a person does not act at all but instead performs someone else's action, and this action strictly speaking falls outside of the

[21]Both the notions of autonomy of a person and of the purpose of a system are extremely problematic notions. As for the latter: the users in the sense of service consumers of a sociotechnical service-providing system view its purpose quite differently from how its owner–user views it, and this latter view differs again sharply depending on whether the owner–user is a state or a private company. As for the former, there are many different accounts of autonomy, typically focusing on different aspects of the concept. My use of autonomy here is in a broad sense characterized by Christman (2009) as "basic autonomy" – "the minimal status of being responsible, independent, and able to speak for oneself" – or in somewhat different words, a person's ability "to act, reflect, and choose on the basis of factors that are somehow her own."

framework where that person is an intentional, let alone a rational agent. Rationality is still an option at a meta-level: it may still be rationally and morally justified to choose to enter such a position, but quite a number of conditions must be satisfied for this. As I see it, whoever takes up this position justifiably[22] must (1) subscribe to the goals of having and operating the "machine" one becomes a component of; (2) trust, and be justified in trusting, its designers, for having adequately designed it to achieve these goals; (3) trust its owner–user, for completing the system and using it so as to achieve these goals; and (4) trust the other operators in the system for acting in accordance with their instructions. This latter trust goes in both directions: an operator must trust higher-level operators for giving him or her adequate instructions and must also trust lower-level operators to faithfully execute any instructions he or she passes on to them.

The facts are, however, that people generally fall short of full trustworthiness in this respect. Are we ever justified in trusting to the level that is required for being justified in accepting an operator position? It is hardly possible to give a general answer to this question. As already stated, not all sociotechnical systems are designed. Many are to some extent ad hoc systems: their composition in terms of subsystems, including their operators, change constantly. But while driving to work in your car over the motorway, you – or rather the instrumental driving system so composed – are implicitly transformed into the object of some system, to be "handled" by its operators in conformity with the system's purpose and operational principle: if road indicators guide you into a particular lane, you submit yourself to the correctness of this "move" by the system. Certainly as the operator of an instrument partial system "for hire," you deliver yourself to whatever system will be created and to whoever will be the owner–user of that system: any taxi driver may be driving a killer to his victim. Epistemic enslavement therefore is not a condition specifically occurring in engineered systems, and computers and car engines are no blacker boxes than colleagues and customers.

What we can say of engineered systems is that detailed knowledge of how the system works or is supposed to work is in principle available and that it is highly relevant that system operators have access to this knowledge. The disaster of Chernobyl was to a large extent generated by a lack of knowledge of basic aspects of the design of system components among its operators. The control rods were designed in such a way – consisting partly of carbon and partly of empty space filled, in the reactor, with water – that lowering them for a maximally elevated position caused an initial increase in reactor activity instead of a decrease, the effect that lowering the control rods is aimed at achieving. This design feature leads to disastrous consequence if not taken into account in the instructions of how the raising and lowering of a reactor's control rods should be handled. In fact a similar situation as caused the explosion of the reactor core in Chernobyl had already

[22]I am ignoring here deviating reasons, for example, the reasons that a spy or a saboteur may have, which may be justifiable.

occurred earlier in a nuclear reactor in Lithuania, with less disastrous consequences.[23] Recorded discussions among operators while in hospital awaiting their deaths showed there was some awareness of this particularity of control rods in place and a vague idea that it might have been causally relevant but they lacked the knowledge that would have enabled them to foresee the size of the effect (Medvedev 1989, pp. 10, 72). Obviously, the lack of a safety culture in the Soviet Union contributed significantly, as the IAEA emphasized ([INSAG-7], pp. 20–22). Still, the operators thought they could rely on the rules that they had been working with successfully for years, and it is unthinkable that they would have acted precisely as they did, had they been aware of the full details of the reactor design.

Valuing the autonomy of operators would then require support in the form of a principle of maximum knowledge of system properties in system operators. In the first instance it applies to the preparation of operators for their tasks. Extending it to maximum on-the-spot knowledge, in the form of a right to question instructions and to receive supporting confirmation and explanation of instructions on request (of the sort that could perhaps have avoided the Überlingen collision; see below), will be extremely controversial. Such a principle goes against the engineering design philosophy of sociotechnical systems and, apart from its effects of the technical efficiency of such systems, may well create more safety hazards than it removes.

What is achieved by a sociotechnical machine (however we conceptualize this) could not be achieved through the actions of completely autonomous intentional agents. Such agents would, for instance, have to *derive* through deliberation on the basis of the totality of their knowledge and the information available to them that the best action to perform if alarm X sounds is indeed to flip switch Y, which, let us assume, is the action prescribed in the instructions for the control room operator. Circumstances do not pause in order to allow agents to go through such deliberations. System design often presupposes that the operator does not look further than the instructions defining his/her role. Typically there is no time to reflect on the appropriateness of an action required, or at least seemingly required, by the operator's manual, in the current circumstances. The smoothness of the system's functioning may even depend on the promptness with which certain hardware states lead to operator actions.

As a consequence, there is often a trade-off between the safety and security of people within the system's reach (customers, operators, and bystanders) and the autonomy of the system's operators. The increasing complexity of such systems, in conjunction with the increasingly complexity and computerization of its hardware

[23]This near disaster had not led, however, to changes in the instructions for this type of reactor. Cf. the entirely analogous situation for ATC-ACAS interference, where the near collision over Japan failed to lead to a review of the inclusion of ACAS in the air traffic regulation system. Or cf. the similar case of the head-on train collision at Ladbroke Grove in 1999, where repeated passings at danger of a particular notorious signal without a resulting collision failed to lead to a redesign of the signal or of the rules for approaching it, until a collision finally did result. In none of these cases was the hazard that ultimately led to the respective accidents brought to the attention of system operators.

components, tends to emphasize the inevitability and even desirability of operator myopia, and an increased acknowledgment of the autonomy of the persons performing operator roles may lead to a decrease in reliability and safety. The choice against autonomy, and for localized operator control only, is followed in any system where strict hierarchical relations exist between the various operator roles. This includes first of all the military but also civilian systems that are to some extent modeled on the hierarchical system of the military. Such systems actually promote the condition of epistemic enslavement for system operators to ensure their general smooth operation from a system perspective, even against counter-indications in individual cases.

In air traffic, for example, as in the military, instructions from air traffic control to aircraft crews are not questioned. In the case of the Überlingen midair collision, the DHL crew could hear air traffic control issuing an instruction to the Russian airplane to descend and could hear the Russian captain accepting this instruction while it was executing its own ACAS RA to descend as well.[24] They did not right away contact ATC about this, however; they only did so after the Swiss air traffic controller repeated its instruction to descend to the Russian Tupolev, and only hesitatingly, but then it was too late. Understandably, Russia added as a comment to the German accident report their opinion that the DHL crew could have done more to prevent the accident ([Überlingen], App. 10). To determine the balance between how often the exercise of operator autonomy can prevent accidents and how often it causes them is difficult, if not impossible. In the most notorious of air traffic accidents, the 1977 Tenerife runway collision, an unsolicited radio message by one of the crews to point out a dangerous situation developing precisely acted as a cause that prevented the other crew from receiving a crucial ATC instruction to stay put ([Tenerife], p. 44).

The responses of system owner–users and regulators to system failures like these (i.e., involving operator actions that are judged erroneous) go in two directions. With respect to interactions between operators and devices (e.g., operator response to an ACAS RA) or interactions between remote operators (e.g., pilot response to ATC instruction or ATC response to pilot reading), the response is a tightening of rules and a corresponding decrease of operator autonomy. With respect to interactions among operators in teams, in contrast, the response is rather an increase of operator autonomy. Approaches like crew resource management are critical of hierarchical forms of organization and emphasize collaboration among equals. Teams of operators are treated as "islands" of purely social systems within a larger sociotechnical environment.

Problems remain in the area that is ambiguous between a regime where interactions are conceived as input–output or stimulus–response and a regime that relies on human intentionality and the capacities of an autonomous agent. A crucial aspect of this is the

[24]The facts about the radio communication are clear from Appendix 3 to the report. It may not have been immediately obvious to the DHL crew that the aircraft receiving these ATC instructions was the one that was approaching them, but it was at least highly likely, likely enough to warrant their interference.

formation of beliefs. To ensure that the right beliefs are formed in the minds of operators is a major design problem. Indeed it is often underestimated how many different models concerning what goes on inside people (especially their minds) we have to rely on in the design of systems with people as components, and how questionable the general applicability of these models is. The Tenerife accident is a case in point: it is perhaps the greatest enigma of this accident what had made the KLM pilot(s)[25] so convinced that the Pan Am Boeing was no longer on the runway. The Dutch commentary on the Spanish investigation report ascribes this to inference on the part of the pilots: the Pan Am Boeing could no longer be on the runway because they had received clearance for takeoff (or so they presumably thought). The emphatic response, however, does not suggest that the conviction was of this inferential sort.

Similarly located in this ambiguous area are cases where operators do not respond to out-of-the-ordinary circumstances or emergencies according to procedures. The crash of Air France Flight 447 into the Atlantic Ocean in June 2009 resulted mainly from a completely inadequate response by the pilot-in-command to the deactivation of the automatic pilot on the loss of signal due to frost from the outboard devices for measuring airspeed. Procedures for this flight situation existed but they were not used nor their existence recalled by either of the two pilots on duty ([Air France 447], p. 175). Likewise a failure to operate according to the specific procedures for the circumstances at hand, this time landing in conditions of severe side wind, led to the crash of Lufthansa flight 2904 in Warsaw in September 1993 ([Lufthansa 2904], p. 42). It is common to attribute such failures to deficiencies in operator training. It is less clear, however, whether training should focus on conditioning behavior or on generating awareness or both, and if the latter, how this should be accomplished. One of the somewhat counterintuitive things that training will have to address is the fact that the direction of technical developments will, increasingly, not only permit but demand minimal operator interference. For the crashes of Air France Flight 447 and Aeroflot Flight 593, postaccident analysis revealed that without the frantic and ongoing attempts of the cockpit crews to steer their way out of the predicament caused by their initial actions, the automated correction and safety mechanisms of the aircraft – Airbus A310 and A330 – would most likely have prevented a fatal flight path.[26] These cases in particular bring out

[25]The Spanish investigation report of the accident ascribes the emphatic answer "jawel" ("yes") in reply to the flight engineer's question whether the Pan Am Boeing had perhaps not yet left the runway to the captain, but according to the Dutch commentary on the report, both pilots gave this reply simultaneously; see [Tenerife], p. 45 and [Tenerife-NL], pp. 46, 63.

[26]This information is presented by experts, with access to the official investigation reports [Air France 447] and [Aeroflot 593], in the respective episodes of the television documentary series *Mayday*, also known as *Air Crash Investigation*, produced by the Canadian firm Cineflix. The episode dedicated to Air France Flight 447 is called "Air France 447: Vanished" and dates from April 2013; the episode on Aeroflot Flight 593 is called "Kid in the Cockpit" and dates from November 2005. The crash of the Airbus leased by Aeroflot was caused by the inability of its Russian crew to handle an unintended and undetected partial disengagement of the autopilot occurring when the captain let his two children, who were also on board, sit in his chair and touch some of the controls.

the tension between the two tendencies of automation and autonomy and the failure to address this tension squarely.

Whatever the views of the potential of continuing automation that will develop in engineering, on the nontechnical side lack of public acceptance and constraining legislation – which may partly reflect the lagging acceptance – will prove formidable obstacles. Sociotechnical systems including human operators will therefore remain a presence for some time to come. Accordingly there is good reason to reflect more on the moral consequences of putting people in that position and the values at stake for them. Autonomy is such a value, and this section has sketched the pressure that the development of sociotechnical systems toward greater reliability and smoothness as well as greater complexity puts upon operator autonomy. In our society, autonomy is closely linked to responsibility, however, and to this key notion I now turn.

## Assigning Responsibility Responsibly

Personal autonomy is considered to be a crucial precondition for the assignment of responsibility, both in a forward-looking sense and in a backward-looking sense associated with liability and blame (cf. Van de Poel 2011 for the distinction). The assignment of responsibility in particular to system operators is taken extremely seriously in our society and treated as a cornerstone of the "license to operate" of the numerous public and private service-providing sociotechnical systems. ICAO regulations, for example, repeatedly stress the ultimate responsibility of pilots for the safety of their aircraft and its passengers and crew regardless of their ever-growing technical embeddedness.[27]

The complexity and scale of current systems have now far outrun the form of control that underlies the assignment of general responsibility. With respect to time scale, for example, the mismatch is obvious in air traffic. The responsibility of pilots for the separation of aircraft, complete with priority rules similar to those that govern road traffic, still underlies international regulation: the ICAO Rules of the Air contain detailed and separate priority rules for head-on approach, convergence, and overtaking ([ICAOAnn2], p. 3–3). One cannot possibly count on this as being of any help to avoid accidents. The rapid increase of midair collisions in the 1940s and 1950s, first between civilian and military aircraft and later between civilian

[27]E.g., [ICAO9863] p. 5–3: "ACAS does not alter or diminish the pilot's basic authority and responsibility to ensure safe flight." [ICAO8168] p. III-3-3-1: "Nothing in the procedures specified [below] shall prevent pilots-in-command from exercising their best judgement and full authority in the choice of the course of action to resolve a traffic conflict." [ICAOAnn2] p. 3–2: "Nothing in these rules shall relieve the pilot-in-command of an aircraft from the responsibility of taking such action [. . .] as will best avert collision."

aircraft mutually, shows that this is not even a recent problem.[28] Nevertheless the investigation report concerning the September 1976 midair collision near Zagreb showed this attitude concerning pilot responsibility: "For the purpose of aircraft collision it is the duty of the crew to look out even though a flight is being made with IFR [Instrument Flight Rules] plan in visual conditions." Since at the time of the collision the weather was fine, "there was nothing to prevent the crew in that respect." However, the investigation report also made clear that the two crews did not notice each other until the collision occurred – one crew never noticed anything since they were killed on impact. Indeed the report itself acknowledges, immediately after referring to the duty of crews, that "[a]t high altitudes and at high speed, particularly in nearly opposite heading it is very difficult to observe another aircraft" ([Zagreb], pp. 32–33). This state of affairs has not changed since. On the contrary, in the most recent midair collision, between a Boeing 737 and an Embraer jet above the Amazon forest in September 2006, neither crew saw the other aircraft coming, although they flew exactly head-on toward each other in clear weather. Since one of the two aircrafts managed to remain airborne, we have first-hand knowledge that the collision itself took such a split second that it was not even experienced as a collision; it was only on the basis of the visible damage to one of their wings that the pilots could infer that they had collided with something ([Amazonas], pp. 235–237).

This is just one particular aspect of a specific type of system. There are more general reasons to question the practice of squarely assigning responsibility to system operators. It is often claimed that it is extremely difficult to make operators accountable in case of accidents or in cases harm is done because their individual actions are just a contribution to the final result, which additionally required the actions of many other operators and the satisfaction of many conditions. This is called the problem of many hands. The problem of many hands, however, was diagnosed and coined for bureaucratic institutions and is typically analyzed and discussed with reference to these (cf. Thompson 1980; Bovens 1998). The situation for sociotechnical systems is actually more complex. Firstly, we are dealing there with the problem of many hands-and-devices rather than the problem of many hands. In one case, the crash during an emergency landing of United Airlines Flight 232 in Sioux City in 1989, the loss of an aircraft and about one third of its passengers, was due to the failure of a turbine fan disk as a result of a microcavity present since manufacture in combination with the failure of the inspection regime to timely detect the cracks that gradually develop from such miniature manufacture defects (see Del Frate et al. 2011, p. 756). Secondly, the hands involved are

[28]It also shows how differently military pilots are treated with respect to civilian ones. The culpable pilot of the military aircraft that collided with a commercial airliner in October 1942 was acquitted. This is typical for accidents involving military operators: as recently as 1998, two American pilots were acquitted who had flown their aircraft low enough to cut through the cables of a cable car in the Italian Dolomites, killing 20 people. This brings out once again that military operators are treated in all respects, including legal aspects, as (and will know they are treated as) true components whose "horizon" is determined by their operator role only. They can "malfunction," but their malfunctioning is exclusively their superiors' concern.

distributed over a much wider range, reaching from the design phase (including the definition of the design task itself!) to the phase of system operation. Some of the difficulties this creates, given the hybrid character of sociotechnical systems, were discussed in the previous section. The bureaucratic systems of government and commercial service-providing systems that were the focus of study until now are not designed in the engineering sense of design; instead they resemble ad hoc systems in that their owner–users can to a large extent be viewed as their designers, with corresponding powers to redesign the system in response to emerging flaws, changing circumstances, or modified objectives.

The intimate connection between operator performance and system structure is acknowledged in Whittingham's claim (2004, p. 254) that "Human error is not inevitable but rather is the inevitable consequence of defective systems," where the systems may be either technological (e.g., the design of control panel displays) or organizational (e.g., the organization of maintenance procedures). The claim strictly speaking does not say that all human errors originate from system defects, only that all defective systems will sooner or later generate human errors. However, for the claim to be substantive, we need a characterization of what a defective system is that is independent of the notion of human error, otherwise the claim does amount to a statement that all human error originates from system defects but at the price of being tautological as a descriptive claim. Instead the tautology could be embraced and we could look upon it as a normative requirement, a design criterion: no system should be allowed to generate human errors or, rather, no system should be allowed to fail because of a human error. Arguably this is how Whittingham takes his claim: he advocates a practice where operators need not fear to be punished for errors committed but are encouraged to report them, so that defects in the system can be remedied.

Whittingham concedes that operators can show "an element of carelessness, inattention, negligence or deliberate violation of rules that must be dealt with," but he seems to downplay the reach of accountability even for such cases, stating no more than that in such cases an operator "deservedly attracts some blame" (pp. 254–255). But the presence of operators with a capacity for autonomy will make the task of designing systems that are immune to operator "errors" a horrendous one. As late as 1993, in a well-established safety culture, reckless and careless pilot behavior could result in the crash of a small airliner with 18 casualties (Tarnow 2000). In a more extreme case in 1999, the pilot of an EgyptAir Boeing 767 steered his aircraft deliberately into the Atlantic Ocean, according to the official US investigation report, taking the 216 other people on board with him.[29] Technical

[29]See [EgyptAir990], pp. 58–65. The Egyptian authorities contested this interpretation, without being able to submit a convincing alternative cause. One can assume they were too embarrassed to admit the facts. This may seem an extraordinary case, but there are at least two similar cases where there is overwhelming evidence that an airliner pilot committed suicide by ditching his aircraft, one from 1997 and one from 2013, with 32 and 103 additional casualties, respectively. There are indications that the still unclarified disappearance of Malaysian Airlines Flight 370 over the Southern Indian Ocean in March 2014, with 239 people on board, may prove to be a fourth example.

safety systems to prevent such accidents will most likely be prohibitive of normal aircraft operation, whereas the intensity of psychological monitoring that would detect such tendencies will be felt as too seriously invading the privacy of operators to be acceptable. Clearly there are limits to the extent that human errors can be reconstructed as system defects. If a system has human operators, it will have autonomous operators, both for technical and for moral reasons.

To avoid the condition of epistemic enslavement as diagnosed by him in certain sociotechnical systems, Van den Hoven (1998) likewise translates his worries in requirements for system design (p. 106): "... the system or epistemic artefact must be designed in such a way as to allow the user to work with it, while retaining his status as a morally autonomous person, who can take his responsibility." In the subsequent discussion, however, the focus shifts to the role of personal autonomy during system operation, with respect to which Van den Hoven proposes the notion of *meta-task responsibility* as an extension of our ordinary notion of responsibility. If a person is to responsibly accept the invitation to execute a particular task – perform a particular operator role – then it is this person's responsibility to see to it that he or she is able to execute this task responsibly. In Van den Hoven's words (p. 108):

> A user A has meta-task responsibility concerning X' implies that A has an obligation to see to it that (1) conditions are such that it is possible to see to it that X is brought about [A's positive task responsibility], (2) conditions are such that it is possible to see to it that no harm is done in seeing to it that X is brought about [A's negative task responsibility].

In order to hold on to a "discourse of responsibility" grounded in personal autonomy, cherished both by philosophers and society, against the equalizing pressure of increasingly complex systems, this puts an additional burden of responsibility on the shoulders of the very same person who has a particular task responsibility. As Rooksby (2009) has argued, this is asking too much. Partially it is very problematic because many operators operate in a hierarchical system: meta-task responsibility would require that subordinate operators check and monitor the actions and inactions of their superiors. Additionally, the required knowledge is vast and may not be easily available; much of it is designer's knowledge. Finally, resources are limited: if you have to explain to your subordinates why you are giving them certain instructions and why it is desirable that they execute them faithfully and why they are justified in executing them faithfully, then it may be that you are left with too little time to attend to your own primary task responsibility.

The notion of meta-task responsibility can indeed play an important role. In my view, however, the perspective should be shifted back to the design context, where according to Van den Hoven originally statement his worries on the ability of operators in complex systems to take responsibility should be addressed. Meta-task responsibility should be seen as implying a constraint not on how persons operating in a sociotechnical system should conceive their responsibility, taking the role they are supposed to play for granted as given by system design, but instead on how system design should conceive of such roles. It is, so I propose, to

be reconceived as a responsibility that people have with respect to the tasks of *others*, tasks that they are responsible for designing. It applies to the design process as (systems) engineering sees it, where operator roles are first conceived and role instructions are first drawn up as part of the grand design of a system module or even the system as a whole. But is also applies to the implementation phase, where owner–users complete the design of a system (while often partly redesigning it) by specifying role instructions and organizing role training and role selection procedures. And it applies, finally, to the operation phase, where ad hoc orders are given by higher-order manager–operators in response to particular circumstances.

This makes the execution of meta-role responsibility a diffuse matter, in that many different parties have a meta-task responsibility with respect to the ability of a single operator to take responsibility for his or her task. But it reflects the hierarchy of the various design regimes that have some amount of control over that task. The proposal should not be read as removing from the person of the operator him- or herself any responsibility for reflection on whether or not to perform what the task seems to require. To the extent that an operator has knowledge of the system and the particular circumstances and has access to the relevant data, and in that sense can be seen as having to some extent designer's knowledge of the operator's role, there is nothing against including the operator in the set of agents who have a meta-task responsibility with respect to their tasks. What matters is to conceive of the responsibility issues from a design perspective, in accord with Van den Hoven's initial emphasis. In view of the discussion of the distinguishing characteristics of sociotechnical systems in the previous section, the implementation of this shift of perspective invites a reorganization of design processes for sociotechnical systems. What that reorganization should involve is, however, not a task that I here undertake.

Against the background of these considerations, the question still has to be asked why we are so keen on assigning responsibility to operators. According to Björnsson (2011, p. 188), "Our interest in holding people responsible is largely an interest in shaping motivational structures – values, preferences, behavioural and emotional habits, etc. – in order to promote or prevent certain kinds of actions or events that we like or dislike." This can be termed an instrumental view on responsibility. This view, however, makes it difficult to uphold the conceptual link with personal autonomy. This is a topic that deserves further discussion. But it seems to me clear that such an instrumental view cannot serve as sufficient justification for assigning a responsibility that will lead a (legal) life of its own once an accident has happened. Pressure from air traffic controllers worldwide led to the Zagreb controller being pardoned after having spent several months in prison, but similar protests did not help the Linate controller, whose conviction was confirmed up till the Corte di Cassazione.[30]

[30]See for the juridical aftermath of both accidents the references given in the Wikipedia articles http://en.wikipedia.org/wiki/1976_Zagreb_mid-air_collision and http://en.wikipedia.org/wiki/Linate_Airport_disaster; see [Linate] for the details of the Linate collision.

Given the current state of engineered sociotechnical systems, how they come to be, how they function, and how they are likely to develop, we – designers and legislators, the latter being, in a democratic society, the entire adult population – owe it to operators to think very carefully what we want to hold them responsible for and what we can justifiably hold them responsible for, and how the responsibility we do assign should be supported by technological means. Additionally, we should think more carefully what the rationale for our society's insistence on apportioning responsibility – our blame culture, as the title of Whittingham's (2004) book suggests – is. Upon reflection we may find that there are good reasons to restrict the responsibility of operators or at least to specify their responsibility in much greater detail than is now customary. Thinking how this can be done in a socially viable and morally justifiable way is a major task for the future.

**Acknowledgments** I thank Luca Del Frate for inspiring conversations and helpful suggestions during the writing of this paper.

## Cross-References

► Design for the Value of Safety
► Design for the Value of Responsibility

## References

### (a) Authored Works

Alexijewitsch S (2013) Secondhand-Zeit: Leben auf den Trümmern des Sozialismus. Hanser, Berlin, Transl. from the Russian original Vremja second-hand: konec krasnogo čeloveka, Moskva: Vremja, 2013

Amalberti R (2001) The paradoxes of almost totally safe transportation systems. Saf Sci 37:109–126

Björnsson G (2011) Joint responsibility without individual control: applying the explanation hypothesis. In: Vincent N, van de Poel I, van den Hoven J (eds) Moral responsibility: beyond free will and determinism. Springer, Dordrecht, pp 181–199

Bovens M (1998) The quest for responsibility: accountability and citizenship in complex organisations. Cambridge University Press, Cambridge

Buede D (2009) The engineering design of systems: models and methods, 2nd edn. Wiley, Hoboken

Checkland P (1981) Systems thinking, systems practice. Wiley, Chichester

Christman J (2009) Autonomy in moral and political philosophy. Stanford Encyclopedia of Philosophy (on-line), substantively revised (first published 2003)

Del Frate L, Zwart SD, Kroes P (2011) Root cause as a U-turn'. Eng Fail Anal 18:747–758

de Weck OL, Roos D, Magee CL (2011) Engineering systems: meeting human needs in a complex technological world. MIT Press, Cambridge, MA
Feld MD (1975) Middle-class society and the rise of military professionalism: the Dutch army, 1589-1609. Armed Forces Soc 1:419–442
Franssen M (2014) Modelling systems in technology as instrumental systems. In: Magnani L (ed) Model-based reasoning in science and technology: theoretical and cognitive issues. Springer, Heidelberg, pp 543–562
Franssen M, Kroes P (2009) Sociotechnical systems. In: Berg Olsen JK, Pedersen SA, Hendricks VF (eds) A companion to the philosophy of technology. Wiley-Blackwell, Malden/Oxford, pp 223–226
Kelly JE (1978) A reappraisal of sociotechnical systems theory. Hum Rel 31:1069–1099
Kleinschmidt H (1999) Using the gun: manual drill and the proliferation of portable firearms. J Mil Hist 63:601–630
Ladkin PB (2004) Causal analysis of the ACAS/TCAS sociotechnical system. In: Cant T (ed) 9th Australian workshop on safety related programmable systems (SCS'04), Brisbane. Conferences in research and practice in information technology, vol 47. unpag
Medvedev G (1989) Chernobyl notebook. JPRS Report no JPRS-UEA-89-034, 23 October 1989. English translation of original Russian publ. by Novy Mir, June 1989
Mitchell CM, Roberts DW (2009) Model-based design of human interaction with complex systems. In: Sage A, Rouse W (eds) Handbook of systems engineering and management. Wiley, Hoboken, pp 837–908
Mumford L (1934) Technics and civilization. Harcourt Brace, New York
Perrow C (1984) Normal accidents. Basic Books, New York
Pritchett AR, Fleming ES, Cleveland WP, Zoetrum JJ, Popescu VM, Thakkar DA (2012a) Pilot interaction with TCAS and Air Traffic Control. In: Smith A (ed) ATACCS'2012, 29–31 May 2012. IRIT Press, London, pp 117–126
Pritchett AR, Fleming ES, Cleveland WP, Zoetrum JJ, Popescu VM, & Thakkar DA (2012b) Pilot's information use during TCAS events, and relationship to compliance to TCAS Resolution Advisories. In: Proceedings of the human factors and ergonomic society 56th annual meeting, Boston 2012, pp 26–30
Rooksby E (2009) How to be a responsible slave: managing the use of expert information systems. Ethics Inf Technol 11:81–90
Sage A, Armstrong J Jr (2000) Introduction to systems engineering. Wiley, New York
Tarnow E (2000) Towards the zero accident goal: assisting the first officer monitor and challenge captain errors. J Aviation/Aerosp Educ Res 10:29–38
Thompson DE (1980) Moral responsibility of public officials: the problem of many hands. Am Polit Sci Rev 74:905–916
van den Hoven MJ (1998) Moral responsibility, public office and information technology. In: Snellen ITM, van den Donk WBHJ (eds) Public administration in an information age: a handbook. IOS Press, Amsterdam, pp 97–111
van de Poel I (2011) The relation between forward-looking and backward-looking responsibility. In: Vincent N, van de Poel I, van den Hoven J (eds) Moral responsibility: beyond free will and determinism. Springer, Dordrecht, pp 37–52
Varfolomeiev T (2007) Soviet rocketry that conquered space. Part 8: successes and failures of a three-stage launcher. Online: http://cosmopark.ru/r7/prig8.htm. Retrieved Jan 2014
Vermaas PE, Kroes P, van de Poel I, Franssen M, Houkes W (2011) A philosophy of technology: from technical artefacts to sociotechnical systems. Morgan & Claypool, San Rafael
Weyer J (2006) Modes of governance of hybrid systems: the mid-air collision at Ueberlingen and the impact of smart technology. Sci Technol Innov Stud 2:127–141
Whittingham RB (2004) The blame machine: why human error causes accidents. Elsevier Butterworth-Heinemann, Oxford/Burlington

## (b) Anonymous Works (Investigation Reports, Regulations)

[AA587] (2004) Aircraft accident report: in-flight separation of vertical stabilizer American Airlines Flight 587 Airbus Industrie A300-605R, N14053 Belle Harbor, New York, 12 Nov 2001. National Transportation Safety Board, Washington, DC

[Aeroflot593] (1995) Akt po rezultatam rassledovanija katastrofy samoleta A310-308 F-OGQS, proissedšej 22 marta 1994 g. v rajone g. Meždurečenska. Departament Vozdušnogo Transporta, Kommisija po Rassledovaniju Aviacionnogo Proissestvija, Moskva. In Russian

[AirFrance447] (2012) Final report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro – Paris. Bureau d'Enquêtes et d'Analyses pour la Sécurité de l'Aviation Civile. Official translation of the original French version

[Amazonas] (2008) Final report A-00X/CENIPA/2008. Comando Aeronáutico, Estado-Maior da Aeronáutica, Centro de Investigação e Prevenção de Acidentes Aeronáuticos, Brasilia. Official translation of the Portuguese version

[Concorde] (2000) Accident on 25 July 2000 at La Patte d'Oie in Gonesse (95) to the Concorde registered F-BTSC operated by Air France. Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile. Official translation of the original French version

[EgyptAir990] (2002) Aircraft accident brief: EgyptAir Flight 990 Boeing 67-366ER, SU-GAP 60 Miles South of Nantucket, 31 Oct 1999. National Transportation Safety Board, Washington, DC

[Eurocontrol] (2002) ACAS II Bulletin, Eurocontrol

[ICAO8168] (2006) Procedures for Air Navigation Services: Aircraft operations. Volume I, Flight procedures, Fifth edn. International Civil Aviation Organization, document 8168

[ICAO9863] (2006) Airborne Collision Avoidance System (ACAS) manual, First edn. International Civil Aviation Organization, document 9863

[ICAOAnn2] (2005) Rules of the air: Annex 2 to the Convention on International Civil Aviation, Tenth edn. International Civil Aviation Organization

[INSAG-7] (1992) INSAG-7; The Chernobyl accident: an updating of INSAG-1. A report by the International Nuclear Safety Advisory Group. International Atomic Energy Agency, Vienna. Safety Series No. 75

[JA907/958] (2002) Aircraft accident investigation report: Japan Airlines Flight 907, Boeing 747-400D, JA8904, Japan Airlines Flight 958, Douglas-DC-10-40, a near midair collision over the sea off Yaizu City, Shizuoka Prefecture, Japan, at about 15:55 JST, January 31, 2001. Aircraft and Railway Accidents Investigation Commission. Official translation of the original Japanese version

[Ladbroke Grove] (2001) The Ladbroke Grove rail inquiry. Part 1 report. The Rt Hon Lord Cullen PC. HSE Books

[Linate] (2004) Final report (as approved by ANSV Board on the 20th of January 2004): accident involved aircraft Boeing MD-87, registration SE-DMA and Cessna 525-A, registration D-IEVX Milano Linate airport, 8 Oct 2001. Agenzia Nazionale per la Sicurezza del Volo, Roma. Official translation of the original Italian version

[Lufthansa2904] (1994) Report on the accident to Airbus A320-211 aircraft in Warsaw on 14 Sep 1993. Unofficial translation of the original Polish report, Państwowa Komisja Badania Wypadków Lotniczych, Warsaw, by Peter Ladkin

[Rules of the Air 2007] (2007) The rules of the air: regulations 2007. Statutory instruments 2007 No. 734; Civil aviation. The Stationery Office Limited, London

[Tenerife] (1978) Joint report K.L.M.-P.A.A. 12.7.1978. Colisión aeronaves Boeing 747 PH-BUF de K.L.M. y Boeing 747 N 737 PA de PanAm en Los Rodeos (Tenerife) el 27 de marzo de 1.977. Ministerio de Transportes y Comunicaciones, Subsecretaría de Aviación Civil, Dirección de Transporte Aereo, Comisión de Accidentes, Madrid

[Tenerife-NL] (1978) Raad voor de Luchtvaart. Netherlands Aviation Safety Board. Final report and comments of the Netherlands Aviation Safety Board of the investigation into the accident with the collision of KLM flight 4805, Boeing 747-206B, PH-BUF, and Pan American flight 1736, Boeing 747–121, N736PA at Tenerife Airport, Spain on 27 Mar 1977. (ICAO Circular 153-AN/56.) October

[Überlingen] (2004) Investigation report AX001-1-2/02 [Report on mid-air collision near Überlingen on 1 July 2002]. Bundesstelle für Flugunfalluntersuchung. Official translation of the original German version

[Zagreb] (1977) British Airways Trident G-AWZT; Inex Adria DC9 YU-AJR: Report on the collision in the Zagreb area, Yugoslavia, on 10 September 1976. Reprint of the report produced by The Yugoslav federal Civil Aviation Administration Aircraft Accident Investigation Commission. Aircraft Accident Report 5/77. Her Majesty's Stationary Office, London

# Design for Values and the Definition, Specification, and Operationalization of Values

Peter Kroes and Ibo van de Poel

## Contents



**Abstract**

This chapter discusses a methodological problem that advocates of design for values have to face. In order to take into account moral values in designing technology, these values have to be operationalized or made measureable; otherwise it will not be possible to evaluate various design options with regard to these values. A comparison of the operationalization of values with the operationalization of physical concepts shows that certain conditions that enable the operationalization of physical concepts in objective measurement procedures are not fulfilled for the operationalization of values. The most significant difference is that physical concepts are embedded in networks of well-tested theories

P. Kroes (✉)
Delft University of Technology, Delft, The Netherlands
e-mail: p.a.kroes@tudelft.nl

I. van de Poel
Department of Ethics and Philosophy of Technology, Delft University of Technology, Delft, The Netherlands
e-mail: i.r.vandepoel@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_11

and operational procedures, which is not the case for moral values. We argue that because of this second-order value judgments play a crucial role in the operationalization of values and that these value judgments seriously undermine any claim that values may be measured in an objective way. The absence of objective measurement of values, however, does not imply that the operationalization and measurement of values in design is arbitrary. In our opinion technical codes and standards may play a major role in coming to a reasonable or justified consensus on how to operationalize and measure moral values in design.

**Keywords**
Design for values • Specification of values • Operationalization of values • Measuring moral values

## Introduction: Design and Value Creation

The design and development of technical artifacts and systems is just one step in a complex process of trying to create value, more in particular to create valuable technical goods and services. Apart from design and development, other steps included in this value creation process are production, sales, after-sales, and use. The different stakeholders involved in this process may have different views on what kind of value is being created. Design engineers may highlight the technical value by stressing technical innovations in and patents on the product, whereas production managers may look at the value created primarily in terms of corporate profits, and sales managers in terms of market position. The end users may appreciate the value of the goods and services in terms of satisfying their needs and reaching their goals; these needs and goals may be very diverse bringing into play various kinds of user values (which, for instance, may be classified as values corresponding to Maslow's five basic human needs). Governmental institutions may look at how the creation, production, and use of technical goods and services enhance public or social values like the health and safety of production workers or users or the privacy of citizens.

So, various kinds of value play a role in the design and production of technical goods and services, including technical, economic, social, and moral ones. Although these various kinds are associated with different phases and stakeholders in the product creation process, there is a strong tendency to take more or all of them into account in the design phase. Technical and economic values and values related to health, safety, and environment play a central role in today's engineering design practice. Advocates of design for values and of socially responsible innovation argue that design engineers should go one step further, namely, that they also should take into account social and moral values in designing technology. This raises the main issue of this chapter, namely, the issue whether it is possible to take such values into account and if so – the possibility of design for values hinges on a positive answer – how this may be achieved. Much progress has already been made with regard to taking into account various kinds of values in engineering design; there are, for

instance, depending on the kind of technical artifact that is being designed, all kinds of norms and standards for values such as health, safety, and environment. Clearly, some of these are highly morally relevant. From this perspective, it seems that the prospects for making progress in taking moral values into account look rather good. So, what are the obstacles, if any, for bringing design for values into practice?

The problem with regard to design for values is a methodological one, which is not specific to taking into account moral values but is of a general design methodological nature. According to design methodology, any functional requirement and any other constraint that the object of design has to satisfy have to be formulated or translated into a list of design specifications. Since any proposed design is going to be evaluated against this list of specifications, specifications have to be formulated as unambiguously as possible, preferably in terms of criteria that may be operationalized in objective measurement procedures. Often the meaning of these criteria and the measurement procedures are fixed in industry or governmental standards. One of the reasons for putting so much effort in standardization is to avoid disagreements about whether or not a particular technical design (technical artifact) satisfies the list of specifications or is "out of specs."

So, if the aim is to make design for values an integral part of engineering design practice, then any constraint imposed on the object of design stemming from social or moral values somehow has to be translated as unambiguously as possible in design specifications, and these in turn have to be operationalized, again as unambiguously as possible, in measurement procedures. In order to explore to what extent the specification and operationalization of moral and social values face problems that are specific for these kinds of values and that may raise doubts about the feasibility of design for values, we will have a closer look at how physical concepts are made measurable. Just as a general definition of, for instance, privacy does not tell what specific constraints a particular object of design has to satisfy in order to protect or enhance the privacy of its users, general definitions of physical concepts such as temperature or mass are not sufficient to put these concepts to "work" in physics. For that it is necessary to operationalize these concepts in terms of measurement procedures. We will take the way concepts are made measurable in physics as our "golden standard" and explore to what extent this standard may be transposed to the specification and operationalization of moral and social values in engineering design.

In order to analyze what it would take to measure moral values in the context of design for values, we proceed in the following way. After a brief look at the philosophical background of the issue of measuring moral values (section "Philosophical Background") and a discussion of some preliminary issues (section "Some Preliminary Issues"), we describe for comparison purposes how the concept of temperature is operationalized in physics (section "Definition and Measurement of Temperature"). This is followed by a discussion of three conditions that a "good" measurement has to satisfy (section "A 'Good' Measurement"). In the next step, we analyze with the help of an example the problems that are encountered in trying to operationalize and measure morally relevant values (section "Value Definition, Specification, and Operationalization: An Example"). In particular we will focus on what is called "specification" of

values and how it relates to the operationalization of values. Thereafter we turn to a discussion of the role of codes and standards in value judgments (section "Codes, Standards, and Value Judgments"). The chapter ends with a brief summary of our main results.

## Philosophical Background

Let us introduce the philosophical background of the methodological issue addressed in this chapter with the help of two quotations. The first quotation is taken from one of Plato's dialogues, *Euthyphro*. In this dialogue Socrates questions Euthyphro about what is the holy and good. Euthyphro professes to know what these notions stand for, and convinced that in this he is doing the good, he is on his way to the Athenian court to accuse his own father of murder! This dialogue contains the following passage (Plato 1973, p. 175):

> SOCRATES: And similarly if we differed on a question of greater length or less, we would take a measurement, and quickly put an end to the dispute.
> EUTHYPHRO: Just that.
> SOCRATES: And so, I fancy, we should have recourse to scales, and settle any question about a heavier or lighter weight?
> EUTHYPHRO: Of course.
> SOCRATES: What sort of thing, then, is it about which we differ, till, unable to arrive at a decision, we might get angry and be enemies to one another? Perhaps you have no answer ready, but listen to me. See if it is not the following – right and wrong, the noble and the base, and good and bad. Are not these the things about which we differ, till, unable to arrive at a decision, we grow hostile, when we do grow hostile, to each other, you and I and everybody else?
> EUTHYPHRO: Yes, Socrates, that is where we differ, on these subjects.

In this dialogue the two disputants come to the agreement that certain differences of opinion may be resolved by measurements, others not. Their examples of problems that may be resolved by measurements are called in modern terms "empirical" problems, problems that may be resolved by observation, whereas the problems that may not so be resolved concern issues about moral values. So, with regard to a certain kind of issues, consensus may be reached (or forced?) by an appeal to measurements. In those cases, it is possible to reveal, so to speak, the true, objective state of affairs in the world simply by observation or performing a measurement. When it comes to differences about moral values, "scales" (methods) for measuring the moral value of something are lacking and so we are "unable to arrive at a decision."

The second quotation stems from The Tanner Lecture on Human Values delivered by Thomas Nagel in 1979 which is entitled *The Limits of Objectivity*. In his Tanner Lecture, Nagel defends the pursuit of objectivity in the domain of ethics. He interprets objectivity as a method of understanding the world; we may arrive at a more objective understanding of the world by stepping back from our own subjective view of the world ("the view from within") and by including ourselves with our subjective view in the world that is to be understood

("the view from without"). However, this way of "objectivizing" the world has its dangers (1979, p. 117):

> So far I have been arguing against scepticism, and in favour of realism and the pursuit of objectivity in the domain of practical reason. But if realism is admitted as a possibility, one is quickly faced with the opposite of the problem of scepticism. This is the problem of over-objectification: the temptation to interpret the objectivity of reasons in too strong and unitary way.
>
> In ethics, as in metaphysics, the allure of objectivity is very great: there is a persistent tendency in both areas to seek a *single, complete* objective account of reality – in the area of value that means a search for the most objective possible account of all reasons for action: the account acceptable from a maximally detached standpoint.

According to Nagel objectivity has its limits and conflicts between objective and subjective reasons for action should be taken seriously in ethical issues. However, from the pursuit of objectivity in ethics, more in particular of objectivizing reasons for action, it appears to follow that we should strive for the most objective possible account of moral evaluations of various options for actions and of states of affairs in the world, since these moral evaluations play an important role in reasons for action. If we assume that there are no a priori methods for doing so, the only way to achieve this, it seems, is to try to introduce objective methods for measuring the moral value (goodness, badness) of actions or states of affairs. If we would succeed in doing so, then, as in the case of length or weight, it would be possible to settle disagreements about moral values with the help of such measurements.

This chapter deals with the methodological question of whether or not or to what extent it is possible to measure objectively moral value (goodness), such that disagreements about moral value (goodness) may be resolved with the help of measurements. Socrates' claim that moral disagreements cannot be settled in an empirical way is widespread, and even the suggestion to explore to what extent moral issues may be settled by empirical measurements may sound strange. After all it is quite common to oppose the domain of the moral (or of the normative in general) to the domain of the empirical: it is taken to be a defining feature of moral (normative) issues that they cannot be resolved empirically. Nevertheless there is, as Nagel points out, the allure of realism and objectivity in the domain of the moral. Indeed, there is a long tradition in philosophy of defending various forms of moral realism, all of which center around the core idea that there are (moral) facts in the world that make moral judgments true or false.[1] If there are such facts, then the question arises why apparently it is not possible to resolve disagreements about moral claims by an appeal to these (moral) facts similar to how disagreements about physical claims may be resolved by an appeal to physical facts. It is not our intention to enter here into a discussion of whether there are such facts, that is, whether moral realism is indeed the case. We will approach the problem of whether measurements may resolve or help in resolving moral issues in a different way. In order to reach a better understanding of the possible role of measurements in

[1] See, for instance, Sayre-McCord (2011).

resolving moral issues, we will analyze in detail the role of measurements in resolving disagreements about physical claims. What is involved in measuring physical quantities and what conditions have to be fulfilled such that measurements can play their role in settling disagreements about physical claims? The answers to these questions will put us in a better position to diagnose the reasons why an appeal to measurements is or may be problematic in the case of moral disagreements.

The methodological issue of measuring moral values (goodness) appears of central importance to any attempt to implement design for values. Somehow, design for values appears to presuppose that at least with regard to some moral values, this is possible. If it would not be possible to measure and compare the moral goodness of various design options, then it seems that the whole idea underlying design for values, namely, that engineers should take moral values into account when designing technical artifacts and systems, loses its rationale. In that case it would be difficult to settle disagreements about the moral value of various design options, since there would be no way of telling which design option is morally better than another.[2] We will argue that in the absence of methods for *objectively* measuring moral values, design for values may still make sense in case there is widespread consensus about which design option is the morally better one; in that case, however, this *intersubjective* consensus is not grounded in objectively measurable features of the design options under consideration. In first instance, however, we are interested in analyzing the conditions that have to be fulfilled so that judgments about the moral goodness of designs may be grounded in objectively measurable features of these designs, just as claims about the physical world may be grounded in objectively measureable features of the world.

## Some Preliminary Issues

Before we enter into a discussion whether moral values may be measured, a number of preliminary remarks are in order. First, of course, there is the issue about the nature of moral values. In the literature there is neither consensus about the meaning of the notion of value in general nor about the meaning of the notion of moral value in particular. For our purposes the following will be sufficient. Examples of moral values of interest within the context of design for values are values such as safety, privacy, sustainability, and accessibility. In contrast to most other values that play a role in engineering design practice (see below), these values

[2]Here we have to point out an important *caveat*. If the overall moral goodness of a design option is not directly measurable but is the aggregated result of the assessment of that design option on various criteria each of which is separately objectively measurable, then in general it will not be possible to compare various design options with regard to their overall moral goodness. In that case the notion of the morally best design option makes no sense. This is due to issues in multiple criteria analysis (see below). What we have in mind here is the assessment of various designs against a "monolithic" moral criterion, that is, a criterion that is not itself the aggregated result of multiple measurable sub-criteria and that may be directly measured in an objective way.

are not instrumental in nature but are pursued primarily (or exclusively) for their own sake because they are intimately related to or an integral aspect of human well-being and human flourishing.[3] So, if we assume that there is some kind of hierarchical ordering of values, design for values deals with values located in the highest regions in a value hierarchy. In the literature these values are often characterized as intrinsic or final values. So, the question we are dealing with is whether values high up in the value hierarchy may be measured objectively.

Second, we are going to use the notion of measurement in a broad sense. Very roughly, a measurement is a representation of relations between certain features of the world in terms of relations between a set of abstract entities. The set of abstract entities is known as the measurement scale. Depending on the measurement scale that is used, a measurement may be classificatory, comparative, or quantitative. The classification of things (states of affairs) in the world in equivalence classes is a measurement on a nominal scale. Suppose that we want to classify persons morally in two types, A and B, and that we have an objective method at our disposal to do so. Then, the classification of a person as of moral type A or B is a measurement on the nominal scale with two measurement values, labeled types A and B (which are labels for two equivalence classes, each class containing all persons who are morally on the same footing). By introducing an ordering on the measurement values of a nominal scale, we get an ordinal scale with regard to which it is possible to perform comparative measurements. In that case, it makes sense to say that a person of type A is morally better (worse) than a person of type B. Suppose that we have at our disposal "moral scales" for comparing the moral goodness of persons, then just as in the case of scales for measuring weights, we would be able to perform a measurement in order to establish which person is morally better (or whether persons are morally on the same footing). If it would also be possible to establish through some kind of measurement *how much* some person is morally better than another, we are entering the domain of quantitative scales (interval and ratio scales). What is important to note is that on our broad notion of measurement, the idea of measuring moral value does not necessarily imply that such a measurement will result in a quantitative value. The claim that design option *A* is morally better than design option *B* may, for instance, amount to an objective comparative measurement of the moral goodness of these design options on just an ordinal scale.

Third, the notion of objectivity in relation to values and measurements stands in need of further clarification. In metaethics there is a long-standing discussion about whether values are real or objective in an ontological sense, that is, whether values are part of the ontological structure of the world. If they are, they are usually taken to be mind independent; if values are real or objective, they are part of the ontological structure of the world independently of the existence of human beings. In that case, values are in Searle's terminology ontologically objective as opposed to ontologically subjective features whose existence is mind dependent (such as the

[3]Note that the fact that a value is pursued for its own sake does not exclude that it may also be pursued for other reasons.

State of France or screwdrivers) (Searle 1995). In this chapter we are not going to address issues about the ontological status of values, nor are we going to make any particular assumptions about their ontological status. We are interested in the question whether knowledge (judgments) about values may be objective and whether measurements of values may form the basis for making objective judgments about values. In Searle's terminology, again, we are interested whether judgments about values may be epistemologically objective, which means that "the facts in the world that make them true or false are independent of anybody's attitudes or feelings about them" (Searle 1995, p. 8). It may be argued that when values are ontologically objective, they are also epistemologically objective – this depends on how the relation between ontology and epistemology is construed – but the reverse appears not to hold. Searle has argued convincingly that objective knowledge of ontologically subjective features of the world is possible: it is, for instance, an objective fact that the State of France exists and that a particular thing is a screwdriver, in spite of the fact that both features of the world are ontologically mind dependent. So, the idea of the epistemological objectivity of moral values is compatible with the idea of their ontological subjectivity. Similarly, it may be argued that the idea that knowledge of moral values is epistemologically subjective is compatible with the idea that moral values are ontologically objective – again this depends on how the relation between ontology and epistemology is construed. Thus, whatever conclusion we may reach with regard to the epistemological objectivity or subjectivity of moral values, it does not commit us to a particular view with regard to the ontological objectivity or subjectivity of moral values.

Fourth, we have to clarify what we mean by the notion of an objective measurement. Intuitively, a measurement is considered to be epistemologically objective if it tells us something about the object on which the measurement is performed, that is, if its outcome is determined only by features of that object, where we leave open whether these features are taken to be ontologically subjective or objective. This intuitive idea and Searle's notion of epistemological objectivity suggest the following necessary condition for an objective measurement: if a measurement is objective, then its outcome does not depend on particular features of the person, the subject, who performs the measurement, such as her or his preferences, points of view, or attitudes. Thus, the outcome of an objective measurement is intersubjectively valid. A measurement of which one of two objects is heavier with the help of scales satisfies this condition; the outcome does not depend on who performs the measurement and is the same for every subject. This is not the case when the measurement is done by comparing the weights of the objects "by hand," for then the outcome may depend on subjective features (wishes, preferences, etc.) of the person who performs the measurement. Thus measuring by hand is not an objective but a subjective measurement method, and in particular cases, it may not be possible to reach an agreement about which object is heavier by this measuring method. Note that it is not the simple fact that a person (subject) performs the measurement by hand that makes this method of measuring subjective; any measurement as an intentional act is performed by a person (subject). What makes this method subjective is the fact that subjective features of the person who performs the measurement may influence its outcome.

This intersubjectivity condition, however, is not strong enough. In order to see why, note that if a measurement satisfies the above condition, this does not imply that that measurement is objective in the above sense. For instance, a systematic error may occur in measurements with the scales due to an error in their construction. Nevertheless, measurements with such scales satisfy the intersubjectivity condition: the outcomes do not depend on subjective features of the person who performs the experiments. What is measured, however, is not some feature of the objects whose weight is being compared but some feature of the system consisting of these objects and the measurement equipment.[4] In order to ensure that a measurement reveals *only* features about the object of the measurement, we also have to require that the measurement outcome is not influenced by features of the measuring device. So for a measurement to be strictly objective, it is necessary that it is transparent (in the sense of not containing any traces) not only with regard to features of the person who performs the measurement but also with regard to features of the measurement equipment.[5] This means that the measurement of epistemologically objective features also does not depend on the *kind* of measurement equipment involved (for instance, temperature may be measured with a mercury thermometer or a thermocouple). Even a measurement that satisfies this stronger condition is not always a *good* measurement, for, as we will see later on in more detail, there may be problems about its validity.

Let us see how our analysis of the notion of (strictly) objective measurement would work out for measurements of moral values. Suppose person *X* has to decide whether person *Y* is a morally good or bad person and, on the basis of the observation of *Y*'s features and behavior (to be interpreted in the broadest sense), concludes that *Y* is a morally good person. On our account of measurement, *X* performs a measurement of the moral value of *Y* on a nominal scale using *himself or herself* as the measuring device. What person *X* as a measuring device does is to represent all the information about *Y*'s features and behavior on a measuring scale with two values, good and bad. Whatever the particular details of how person *X* as a measuring device does this, if the outcome of the measurement and the measurement method is objective, then the outcome is not influenced by *X*'s preferences, points of view, attitudes, etc. In that case, anybody making the same observations of *Y*'s features and behavior and making use of the same measurement device (i.e., the same way of representing information about features and behavior of a person on the measurement scale as used by *X*) will come to the same measurement result. As in the case of the weighing scales, this does not exclude the possibility that the

[4]Of course, it is possible to correct for systematic errors of measuring devices such that the corrected outcome is determined only by features of the object(s) on which the measurement is performed (for instance, one may correct the outcome of a measurement with scales for the fact that the arms of the scales are not of the same length). In our opinion, however, we are then dealing with two different kinds of measurement methods, the original one and one with a correction procedure.

[5]For an interesting discussion of the notion of transparency of experimental equipment, including measurement devices, see Lelas (1993).

measurement method itself does influence the outcome. *X*'s measurement device may lead to systematic errors in the outcome.

So far we have been discussing necessary conditions for (strictly) objective measurements. We will leave it an open matter whether these conditions may also be considered to be sufficient. Suffice it here to remark the following. Suppose that there is intersubjective agreement (consensus) about person *X* being morally better than person *Y* (even *Y* agrees!). Does this mean that it is an epistemologically objective feature of *X* that (s)he is morally better than *Y*? That depends on the nature of this consensus. If it is the result of everybody applying the same measurement method, along the lines sketched above, then this case looks similar to the one in which scales are used to reach consensus about which one of two objects is heavier. It is generally assumed that in the latter case, given the absence of systematic errors, we are dealing with an objective measurement and with epistemologically objective features. Since there seem to be no significant differences between the two cases, we do not see any reason to question the epistemological objectivity of the moral features and of the objectivity of comparatively measuring moral goodness. This conclusion, however, is based on a big "if," namely, that the consensus about the moral evaluation is based on the use of a *common* measurement method.

As the quotation from Plato suggests, we do not have at our disposal a common measurement method for moral goodness. So if de facto there is consensus about *X* being morally better than *Y*, then this consensus is not forced, so to speak, by the use of a common measurement procedure. Everyone performs his or her own measurement, often without a clear insight into the details of their measurement method, but the outcome of all measurements is nevertheless the same. In that case the inference from consensus (intersubjective agreement) to epistemologically objectivity becomes much more problematic. The main reason for this is the relation between the meaning of the notion of moral goodness and the way it is measured or operationalized. Suppose that the consensus is the result of the use of various measurement procedures for moral goodness. Schematically, two different situations may be distinguished. In the first situation, all these measurement procedures correspond to different operationalizations of the same notion of moral goodness, similarly to the different operationalizations and measurement methods of, for instance, the notion of temperature in physics. In the second situation, different notions of moral goodness are at play, each with their own operationalization and corresponding measurement method. In the first case, there is no reason to doubt the inference from intersubjectivity to epistemological objectivity. But of course it will be necessary to underpin the claim that the intersubjectivity with regard to some particular moral issue is indeed grounded in different ways of operationalizing and measuring the *same* concept. How is that to be done? As we will see shortly for the concept of temperature, physics offers detailed conceptual frameworks and theories about what temperature is and how it may be measured to support the claim that the various measurement methods for temperature are all operationalizations of the same concept. Nothing that comes near to this exists in the field of ethics and with regard to the concept of moral goodness. That is one of the main reasons why in this domain it is so difficult to make a strong case for the

inference from intersubjectivity to epistemological objectivity. In the second case, any inference from intersubjectivity to epistemological objectivity appears to be out of the question, since the concept of moral goodness has various meanings and various things are being measured depending on the meaning attached to the notion of moral goodness. What on the face of it seems to be consensus about a moral judgment ("person *X* is morally better than person *Y*") is then on closer inspection not a consensus at all, since there is no agreement about the meaning of this judgment.

A final preliminary remark concerns issues about aggregation of values and value (in)commensurability. Consider a situation in which the moral goodness of a person is operationalized in such way that it is measured in terms of different criteria (such as honesty, justice, and altruism). Then the question arises how evaluations of a person against these various criteria separately may be aggregated into an overall evaluation of that person's moral goodness. In general, such an aggregation is necessary in order to be able to compare the moral goodness of different persons. This aggregation problem involves issues about whether different values may be compared to each other or not. In our discussion of measuring values, we will run into issues about aggregation of values and value (in)commensurability, and although we recognize the relevance and importance of these issues, we will not discuss them in any depth (but see chapter "▶ Conflicting Values in Design for Values" in this handbook).

## Definition and Measurement of Temperature

It has taken physicists and engineers centuries to develop a clear notion of temperature, of its unit and scale, and methods of how to measure it quantitatively. Moreover, it seems that this development has not yet reached an endpoint and is still on its way; for instance, the latest revision of the definition of the International Temperature Scale dates back to 1990 (see below). Without doing any justice to the complex history of the notion of temperature and to the complexity of the modern notion itself, the following remarks suffice for our purposes (for more details, see Chang (2004)).

From a phenomenological point of view, the notion of temperature has always been associated with the distinction between warm and cold and with the notion of heat and has been taken to be some kind of measure of the hotness or coldness of things. Within physics, it was only after a clear distinction between intensity of heat and quantity of heat was made and the idea of heat as a form of fluid ("caloric") was given up that the modern notion of temperature established itself during the nineteenth century. Before that time various reliable ways to measure temperature and various temperature scales and units (Celsius and Fahrenheit) had already been introduced. According to most definitions of temperature to be found in present day introductory physics textbooks, the temperature of an object is a measure of the disorderly (random) motion of the particles of which it is made up, more in particular of their mean kinetic energy. For an ideal gas, the temperature is defined more precisely as a measure proportional to the mean translational kinetic energy of its particles.

Physics, however, has much more to say about the notion of temperature than there is to be found in introductory textbooks. Actually, various notions of temperature are in use in physics. The interpretation of temperature in terms of mean kinetic energy works fine at the macroscopic level for particular kinds of systems (composed of atoms and/or molecules). But physicists have developed notions of temperature for other kinds of systems, including systems consisting of photons (electromagnetic radiation; they speak about the temperature of "blackbody radiation") and spins. Apart from that, thermodynamics offers the following definition of absolute thermodynamic temperature in terms of energy and entropy:

$$T = \frac{dq_{\text{rev}}}{dS}$$

which is independent of the particular physical makeup of the system under consideration. This notion of temperature is only well defined for systems that exchange energy with their environment in a reversible way and thus are in equilibrium with their environment. Just as in the case of an ideal gas, absolute thermodynamic temperature is defined with reference to an ideal kind of system since in practice heat exchange is not reversible.

All in all, the situation with regard to the notion of temperature in physics is that it is defined in different ways in different theoretical frameworks, but it may be shown, theoretically as well as empirically, that these various notions of temperature hang together and because of this it is assumed that they all refer to one and the same physical quantity. This is reflected in the fact that measurements of all of these various theoretical notions of temperature share a common temperature scale, namely, the International Temperature Scale of 1990 (T90) referred to above. This scale is defined in the following way[6]:

> Between 0.65 K and 5.0 K T90 is defined in terms of the vapour-pressure temperature relations of 3He and 4He.
> Between 3.0 K and the triple point of neon (24.5561 K) T90 is defined by means of a helium gas thermometer calibrated at three experimentally realizable temperatures having assigned numerical values (defining fixed points) and using specified interpolation procedures.
> Between the triple point of equilibrium hydrogen (13.8033 K) and the freezing point of silver (961.78 °C) T90 is defined by means of platinum resistance thermometers calibrated at specified sets of defining fixed points and using specified interpolation procedures.
> Above the freezing point of silver (961.78 °C) T90 is defined in terms of a defining fixed point and the Planck radiation law.

The unit of this temperature scale, the kelvin, is defined as the fraction 1/273.16 of the temperature of the triple point of water.[7] This does not concern ordinary

[6]See http://www.its-90.com/its-90p3.html.

[7]See http://physics.nist.gov/cuu/Units/current.html.

water; for measurement purposes a standardized form of water with a specific isotopic composition is used known as Vienna Standard Mean Ocean Water.[8]

The technical details of the definition of the temperature scale and its unit do not concern us. They are presented here because they illustrate an important point, namely, that not one measurement procedure is used for defining the whole temperature scale in one stroke; apparently it is considered better to refer to different measurement methods in different regions of the temperature scale. This should not come as a surprise. It is no use trying to measure temperatures in the region of 10,000 K with a mercury thermometer. In general, when we want to measure the temperature of something, it is necessary not only to specify the relevant temperature range but also other physical characteristics of that something. For instance, even if the temperature of a tiny drop of water lies within a temperature range of a mercury thermometer, it does not make sense to use that measurement method, because, as we will see shortly, it will not lead to valid measurements. Thus, *specification* of the conditions under which a physical quantity such as temperature is to be measured (the kinds of system and the temperature range involved) is an important step in making that quantity measurable.

The definition of the temperature scale and unit illustrates yet another point. They do not define the *meaning* of the notion of temperature in the sense of what kind of physical quantity is being measured and of what the kelvin is the unit measure. What is defined is *how* a certain quantity called "temperature" is to be measured by specifying for various regions of the temperature scale a specific measurement procedure against which other measurement methods that may be used in that region are to be calibrated. If these other measurement methods lead to results which are coherent with the definition of the standard in that region, then they are supposed to measure the same physical quantity. There is no reference to notions like mean kinetic energy or other theoretical notions that play a role in the various theoretical definitions of temperature.

The foregoing does not mean that theory plays no role in measuring temperature. On the contrary, these definitions are the outcome of numerous developments in theoretical and experimental physics, and the idea that these various methods all measure the same physical quantity is anchored deeply in theoretical as well as empirical considerations. According to Chang (2004, p. 212 ff), the modern accurate methods for measuring temperature are the outcome of a long process of successful convergence of iterative attempts to improve our theoretical conceptions of and measurement methods for temperature. This is not to be interpreted as a convergence toward the measurement of the "true" value of the absolute thermodynamic temperature of a physical system. For Chang (2004, p. 207) an unoperationalized abstract concept like absolute thermodynamic temperature "does not correspond to anything definite in the realm of physical operations, which is where values of physical quantities belong." The true or real value of the

[8]See http://en.wikipedia.org/wiki/Vienna_Standard_Mean_Ocean_Water and http://www.bipm.org/en/si/si_brochure/chapter2/2-1/kelvin.html.

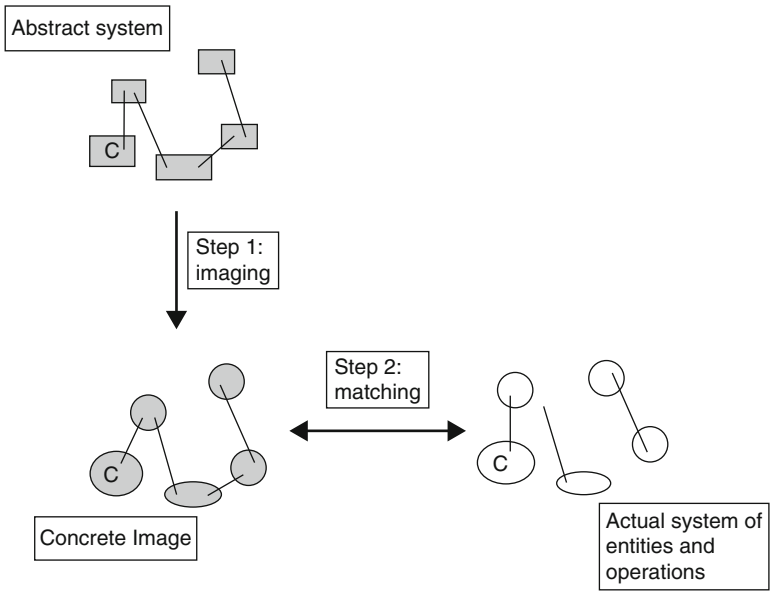



**Fig. 1** Chang's two-step view of operationalization

temperature of a physical system is defined by the measurement method that is the outcome of successful convergence of repeated attempts to operationalize the abstract notion of temperature; in other words, the real value is constituted by a successful operationalization (Chang 2004, p. 217).

So in general the problem of measuring a physical quantity is a problem of how to connect abstract (theoretical) concepts of this quantity to real physical systems and operations on these systems, that is, how to make "contact between thinking and doing" (Chang 2004, p. 197). Abstract physical concepts have, so to speak, no grip on real physical systems; as we will see below, the same applies to abstract moral values and technical designs. As an alternative to the widespread idea that the operationalization of abstract concepts may be achieved through correspondence rules that directly connect these concepts to empirical terms, Chang (2004, pp. 206–207) offers a two-step view on how operationalization of abstract concepts may be achieved (see Fig. 1). The first step consists in finding a concrete image for a system of abstract concepts. A concrete image is not a real physical system but an "imagined system, a conception, consisting of conceivable physical entities and operations, but not actual ones" (Chang 2004, p. 206). The conception of an ideal gas is an example of a concrete image; it is an imagined system whose behavior is governed by an abstract system (the concepts and laws of thermodynamics). The second step consists in finding a match between the imagined system and actual, real physical systems. If no exact match is possible but real physical systems may be configured such that they approach the imagined system, then the concrete image may be characterized as an idealized system. According to Chang the notion of a

valid operationalization has to be interpreted in terms of what he calls a "good correspondence" between the abstract system and concrete image and between the latter and actual physical systems. Unfortunately, he leaves us more or less in the dark about the details of what constitutes a good correspondence. He claims that it is not a one-dimensional notion and that any assessment of the validity of an operationalization calls for a complex judgment.

Finally, the history of the notion of temperature in physics shows that reliable ways of measuring a physical quantity may exist in spite of the fact that the nature of that physical quantity, that is, the meaning of the corresponding notion, may still be under dispute. It shows that, historically, measurement procedures are not always the end result of specifying and operationalizing a well-defined notion. On the contrary, as the history of the notion of temperature shows reliable measurement, methods often play a key role in arriving at consensus about the nature of what is being measured. It may be questioned whether generally speaking the definition of a concept always comes prior to its specification and operationalization. For example, from an operationalist's point of view, the measurement procedure appears to be conceptually prior since the meaning of a concept is identified with its measurement procedure.

The operationalization of concepts plays a key role not only in science but also in engineering, especially in the context of drawing up design requirements. Vincenti (1990, Chap. 3) contains a detailed description of how engineers in the early days of powered flight wrestled with and finally solved the problem of defining and operationalizing the notion of "flying quality" of airplanes. He describes how they set up a research program to deal with issues about what it means for an airplane to have good flying qualities and how those might be measured. Finally, they were successful and were able to specify measurable (testable) design requirements for flying qualities for airplanes. When attempts to operationalize concepts are successful, it is possible to settle disputes about those concepts by an appeal to measurement (testing).

Before we turn to a discussion of the measurement of moral values, it will be necessary to pause for a moment on the question of what makes a measurement a "good" measurement.

## A "Good" Measurement

When is a measurement method a good way of measuring something, that is, when may we consider the outcome of measuring something to be a good measurement, assuming that the measurement method has been applied correctly? If we take the notion of measurement in our broad sense, then this is a crucial question for *any* attempt to settle whatever issue in an empirical way, since then even the categorization of some observable event or object into a particular class involves a measurement. If an appeal to measurement is made in the context of settling a disagreement, then a "good measurement" must satisfy certain necessary conditions in order to assure that the measurement will make it possible to settle the dispute in

an unambiguous way, that is, to "force" consensus among rational disputants. Ideally, these necessary conditions together are sufficient and then define the notion of a good measurement. At least the following three questions appear to be important with regard to the notion of a good measurement in the context of a dispute:

1. Validity: Does what is measured correspond to what one intended to measure in order to settle the dispute?
2. Reproducibility: Is the outcome of the measurement independent of the person who performs it?
3. Accuracy: Is the outcome of the measurement accurate, that is, to what extent does it correspond to the "real" state of affairs in the world (to the "real" value in the case of quantitative measurements)?[9]

If the answer to one of these questions is negative or under dispute, then clearly an appeal to measurement will not settle the issue, because the (outcome of the) measurement itself may become the object of dispute. In general, any good measurement has to be valid, reproducible, and accurate. We will briefly discuss each of these features and illustrate their relevance with the help of measurement methods for temperature.

A measurement method is called valid if it measures what it is supposed or claimed to measure. Various notions of validity are in use (especially in the social sciences). Here we concentrate on what is often called *construct* validity of a measurement method: a measurement method is construct valid if it measures the theoretical (abstract) notion it is intended to measure. For instance, if temperature is theoretically defined in terms of the mean kinetic energy of particles, then the measurement of the temperature of a cup of tea with a mercury thermometer is a valid measurement method (on the assumption that the measurement procedure is correctly executed). The same is true when a thermocouple is used. What is measured

[9]In the literature accuracy is often taken to be part of the notion of validity; see, for instance, Carmines and Zeller (1979). Conceptually, however, a distinction can be made between the questions whether a measurement method measures the intended quantity (e.g., temperature) of the system under consideration or not and if so, how accurate the measurement method is. That is the reason why we prefer to distinguish between validity and accuracy. In specific cases, however, it may be difficult to distinguish between accuracy and construct validity (see below for the notion of construct validity). Consider the case in which someone tries to measure the temperature of a liquid with a mercury thermometer. Suppose that the amount of heat transferred from the liquid to the tip of the thermometer is small compared to the total amount of heat in the liquid. Then, the smaller the heat transfer from the liquid to the tip of thermometer is, the more accurate the measurement will be. But now suppose that the amount of heat transferred becomes more or less equal to or less than the amount of heat in the liquid (e.g., someone tries to measure the temperature of a drop of water with an ordinary mercury thermometer). Then the measurement becomes less accurate or even construct invalid, for in the extreme case of a very small amount of liquid compared to the mercury in the tip of the thermometer, one no longer measures the temperature of the drop of liquid but of the ambient temperature in which the thermometer is kept. This example shows that under certain conditions, very inaccurate measurements may become construct invalid.

is the intended physical quantity, namely, the temperature of the cup of tea. Construct validity presupposes that there is a "theoretical network that surrounds the concept" and is "concerned with the extent to which a particular measure relates to other measures consistent with theoretically derived hypotheses concerning the concepts (or constructs) that are being measured" (Carmines and Zeller 1979, p. 23). As we observed in the previous section, this is indeed the case for the physical concept of temperature. It is embedded in a whole network of theories which explain why measurements of temperature based on the expansion of liquids or on the generation of a voltage difference over a boundary layer between two metals lead to consistent results in situations where both measurement methods can be applied.

To illustrate the importance of the theoretical network for assessing the construct validity of a measurement method, consider the situation in which the temperature of a cup of tea is measured by simply measuring the volume of the tea contained in the cup. This is clearly a nonvalid measurement; its outcome is a particular volume and as such has no relation to temperature at all. The theoretical network in which the notions of temperature and volume of a fluid are embedded does not allow interpreting this particular volume in terms of the temperature of the tea. However, measurement of changes in the volume of the tea may be taken as a valid measurement of changes in its temperature, since there is a theoretically and empirically grounded relation between changes in the volume of a fluid and changes in its temperature. So, changes in temperature may be measured validly by changes in volume (or changes in voltage, or changes in radiation spectrum, etc.) on condition that a suitable theoretical background is in place.

On top of being construct valid, a good measurement has to be reproducible: the outcome of a good measurement may not depend on specific features of the person who performs the measurement (see also the discussion in section "Some Preliminary Issues"). As we have stressed in the foregoing, reproducibility is intended to safeguard the objectivity of the measurement outcome.[10] In equipment for measuring temperature, automatically the person who performs the measurement is more or less eliminated (her role is reduced to switching on the measuring apparatus or even not that; think of the thermostat measuring the temperature in a room) and then reproducibility is usually not an issue. But when temperature is measured by means of the human body (e.g., by hand), then reproducibility may become a real issue; for instance, two persons may systematically, that is, after repeated measurements, disagree about which one of two objects is warmer than the other.

Finally, there is the notion of the accuracy of a measurement method. It is usually interpreted in terms of the extent to which the measured value corresponds to the real value. For a measurement to be accurate, it is necessary that the outcome is not influenced by features of the measuring instruments (see the discussion about

[10]Reproducibility does not mean that if the measurement is repeated by the same person, exactly the same result will come out. Due to (random) measurement errors, the outcomes will be distributed according to a certain probability function. Reproducibility requires that this probability function over the outcomes is the same when the measurement is performed by another person.

the transparency of the measurement equipment in section "Some Preliminary Issues"). The more accurate a measurement procedure is, the higher the chance will be that the outcome of a measurement is closer to the real value. The notion of real value, however, has to be interpreted with care. It is not simply the value that corresponds to the objective state of affairs "out there." It would be a mistake, as Chang has pointed out with regard to the notion of temperature, to think that the objective state of affairs includes a definite value of the absolute thermodynamic temperature for a system under consideration. Without an operationalization of that concept, the system has no absolute thermodynamic temperature. The real value, we propose, may be interpreted as the value that will be measured in the long run when the successful convergence of corrective epistemic iterations in the field of temperature physics (see Chang (2004, pp. 44–48) has come to an end. This real value as such is not part of the objective states of affairs but is constituted by that state of affairs in combination with the operationalization procedures that will be adopted in the long run.

In the following section, we will explore the extent to which these ideas about operationalization of concepts and about good measurements taken mainly from physics may be applied when it comes to measuring moral values in the context of design for values.

## Value Definition, Specification, and Operationalization: An Example

To see how values can be operationalized in design for values and whether this can result in a "good" measurement of values, we will look at an example. The example we will consider is the "design" of a new coolant for household refrigerators in the 1990s (van de Poel 2001).

Before the 1990s, CFC 12 was the commonly used coolant for household refrigerators of the vapor compression type. It had come into common use in the 1930s when Thomas Midgley invented the CFCs. After 1970, however, the use of CFCs came under increasing pressure due to their contribution to degradation of the ozone layer. In 1987, the Montreal Treaty called for a substantive reduction in the use of CFCs. International conferences following the Montreal Treaty recommended yet tougher measures, and during the 1990s many Western countries decided to ban CFCs.

In the search for alternatives to CFC 12, three values played a key role: environmental sustainability, health, and safety. Each of these values has moral significance, and the better an alternative coolant scores on each of these values, the better it is from a moral point of view. So, taken together these three values may be said to determine the "moral goodness" of different potential alternatives. The values were operationalized in a two-step process (see Fig. 2). In a first step, the values were associated with certain evaluation criteria that are more concrete and specific than the values. This step is somewhat comparable with step 1 in Fig. 1. This step associates a more concrete image with an abstract concept. Similarly, evaluation criteria can be seen as the more concrete image of abstract values.

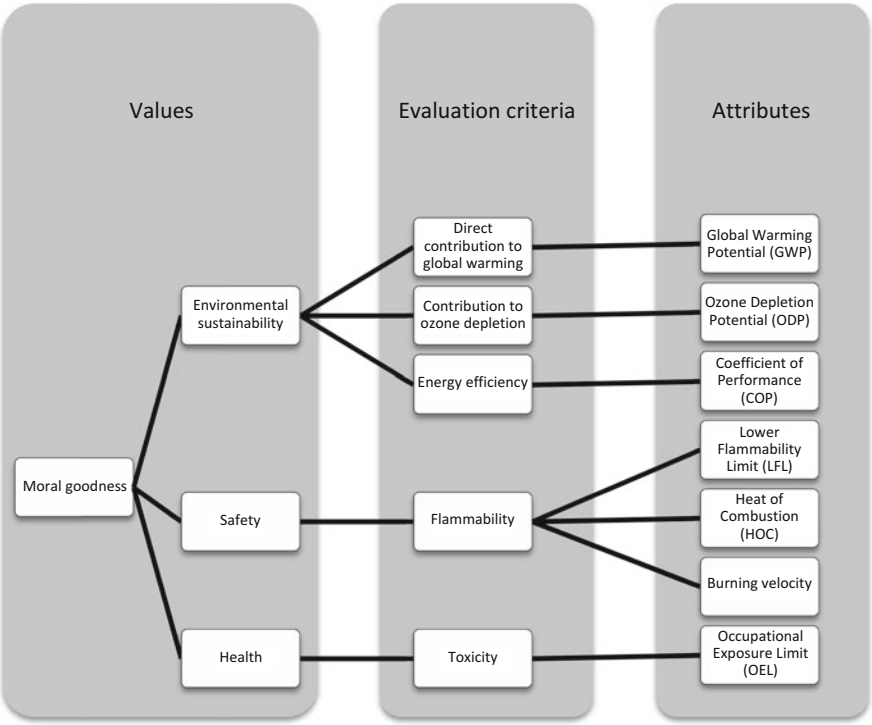



**Fig. 2** Operationalization of the moral goodness of alternative coolants

However, the evaluation criteria are not directly measurable. To make them measurable, they have to be matched with attributes that can be readily measured and for which (standard) measurement methods exist. This step is comparable with the second step in Fig. 2 in which the concrete image (evaluation criteria in our case) is matched with actual entities and processes (attributes in our case).

It is important to note that both steps in Fig. 2, from values to evaluation criteria and from evaluation criteria to attributes, involve value judgments that have to be carefully distinguished from value judgments about the moral goodness of a particular refrigerant. The latter may be called first-order (or object-level) value judgments; they are judgments about the moral goodness of a particular refrigerant (design). The former are second-order (or meta-level) value judgments; they are value judgments involved in operationalizing moral values in a specific way. The role of second-order value judgments in choosing attributes is clearly pointed out by Keeney and Gregory (2005). They state that good attributes for measuring evaluation criteria[11] have to satisfy a number of conditions under which they list comprehensiveness. An attribute (measure) is comprehensive when its levels cover

[11]They speak of objectives but these are similar to what we call evaluation criteria.

all possible forms of achieving the evaluation criterion and any value judgment expressed in the attribute is reasonable. They give the following example (Keeney and Gregory 2005, p. 4):

> Comprehensiveness . . . requires that one consider the appropriateness of value judgments embedded in attributes. Whenever an attribute involves counting, such as the number of fatalities, there is the assumption that each of the items counted is equivalent. With the number of fatal heart attacks, there is a built-in assumption that a fatal heart attack for a 45-year-old is equivalent to a fatal heart attack for a 90-year-old. Is this a reasonable value judgment for a particular decision? There is not a right or wrong answer, but it is an issue that should be considered in selecting the attribute.

According to Keeney (1992, p. 100) "the assignment of attributes to measure objectives always requires values judgments." This is in fact also visible in the refrigerant case. One example is the use of global warming potential (GWP) as an attribute for the evaluation criterion "direct contribution to global warming." GWP can be measured on different so-called integrated time horizons, i.e., different time spans over which the contribution of a certain substance to global warming is integrated. The choice of a specific time horizon involves a second-order value judgment, because it reflects a judgment about what are appropriate time horizons, for example, in the light of considerations of intergenerational justice that are part of the first-order value of "environmental sustainability," of which the GWP is an operationalization.

Second-order value judgments are also involved in the first step of Fig. 2, leading from the overall value of moral goodness to more specific values and from there to evaluation criteria. This is related to the fact that the operationalization of values in design is always context dependent. This can be clearly seen in Fig. 2: it would be absurd to claim that the attributes mentioned in Fig. 2 measure "moral goodness" *in general*! Rather, they are meant to measure "moral goodness" in a very specific situation, namely, the "moral goodness" of potential alternative coolants to be used in household refrigerators. The choice of which specific values and evaluation criteria should be taken into account in a specific context itself involves (second-order) value judgments.

A closer look at the various "translations" made in Fig. 2 reveals that context and second-order value judgments play a different role in the various steps involved in the operationalization of values:

- The association of certain values (like safety, health, and environmental sustainability) with "moral goodness" is context dependent. It depends on second-order *value judgments* on which values are affected by a design and should be taken into account in the design process.
- The definition (and conceptualization) of values (like safety, health, and environmental sustainability) is largely context independent as there are general definitions available (e.g., in moral philosophy or in law). However, there may be *lack of consensus* on how to define (and conceptualize) these values.
- The specification of values in terms of evaluation criteria is context dependent as it depends on the specific product (or class of products) designed. Similarly the selection of certain attributes to measure an evaluation criterion is context dependent. Both steps also involve second-order *value judgments*.

- Measurement methods for attributes in design will often (but not always) be context independent as for many relevant attributes, general measurement methods are available. Here value judgments and lack of consensus play a minor role.

From the foregoing we may draw the conclusion that the operationalization of moral values can never be "objective" in the sense that the operationalization can be rationally derived from the meaning of the value without intermediary second-order value judgments. Prima facie a comparison of Figs. 1 and 2 may leave the impression that moral values can be operationalized in a two-step way similar to how physical quantities like temperature are operationalized. But a closer look reveals a significant difference. Because of the role of second-order value judgments in the operationalization of values, it may always be questioned whether a particular operationalization of a value will result in a "good" measurement of that value.[12] In section "A 'Good' Measurement," we distinguished three considerations in judging measurements: the *validity*, the *reproducibility*, and the *accuracy* of a measurement. The attributes mentioned in Fig. 2 can be measured in a reproducible and accurate way.[13] It is, however, far less clear that they result in a *valid* measurement of the values and ultimately of moral goodness. Do the attributes together indeed measure the "moral goodness" of a certain alternative refrigerant? The issue here is one of construct validity. We have seen that in the case of the measurement of temperature, construct validity is achieved (or at least enabled) by a network of theories and measurement procedures. The crucial difference between the measurement of physical quantities and the measurement of values is that in the case of the measurement of values, we lack such a network of theories to guide the choice of second-order value judgments; as a result these second-order value judgments seriously undermine the construct validity of any measurement of values.

## Codes, Standards, and Value Judgments

Second-order value judgments may be indispensable in operationalizing and measuring moral values in design, but that does not mean that such value judgments need be arbitrary. One may ask whether there is anything in the case of design and moral values that may take the place that the network of theories plays in

[12]Note that, similar to second-order value judgments in the case of the operationalization of values, second-order epistemic value judgments are important in the operationalization of physical concepts. In general, however, these second-order epistemic value judgments appear not to undermine the construct validity of measurement procedures for physical concepts; we will not enter here into a discussion of why this is the case.

[13]At least reproducibility and accuracy are not fundamentally more problematic than in the case of physical quantities because the attributes are, at least in this case, all physical quantities. It may not always be possible to operationalize values in terms of physical quantities, and in such cases reproducibility and accuracy are more of an issue. But even, then, we would argue the real issue is validity.

**Table 1** Refrigerant safety group classification (Based on ASHRAE 2013b, Fig. 6.1.4)

| | Lower toxicity | Higher toxicity |
|---|---|---|
| No flame propagation | A1 | A2 |
| Lower flammability | B1 | B2 |
| Higher flammability | A3 | B3 |

operationalizing physical concepts. One option here may be so-called technical codes and standards. Technical codes are legal requirements that are enforced by a governmental body to protect safety, health, and other relevant values (Hunter 1997). Standards are usually not legally binding, but they might be designated as a possible, and sometimes even mandatory, way to meet a code; they may also play a role in business contracts and are sometimes seen as describing good design practice, and as such they may also play a role in litigation.

Codes and standards often play a prime role in the operationalization and measurement of moral values in design. In the coolants' case, for example, the specification of the values of safety and health in terms of flammability and toxicity, and the attributes matched to these evaluation criteria, was directly based on technical codes and standards. In this case the ANSI/ASHRAE standards 15 and 34 played a major role.[14] Standard 34 ("Designation and Safety Classification of Refrigerants") says the following:

> 6.1 Refrigerants shall be classified into safety groups according to the following criteria.
> 6.1.1 Classification. The safety classification shall consist of two alphanumeric characters (e.g., "A2" or "B1"). The capital letter indicates the toxicity as determined by Section 6.1.2; the arabic numeral denotes the flammability as determined by Section 6.1.3.
> 6.1.2 Toxicity Classification. Refrigerants shall be assigned to one of two classes—A or B—based on allowable exposure: Class A refrigerants have an OEL of 400 ppm or greater. Class B refrigerants have an OEL of less than 400 ppm.
> 6.1.3 Flammability Classification. Refrigerants shall be assigned to one of three classes (1, 2, or 3) and one optional subclass (2 L) based on lower flammability limit testing, heat of combustion, and the optional burning velocity measurement. (ASHRAE 2013b, p. 14)

ASHRAE standard 34, then, results in six safety classes for refrigerants as indicated in Table 1:

Standard 15 (Safety Standard for Refrigeration Systems) prescribes which of these safety classes are allowed, and in what maximum amounts, for different kinds of refrigerating applications. The initial versions of standard 15 allowed unlimited use of A1 refrigerants in household refrigerators, forbade the use of A3 and B3 refrigerants, and set limits to all other categories, in this way guaranteeing a certain level of safety and health protection (ASHRAE 1994).

It should be noted that there are basically two kinds of standards: (1) standards for measurement and classification, like ANSI/ASHRAE standard 34 and, for example, the European Standard IEC 60079-20-1, and (2) standards setting (minimal) design requirements (or certain performances) like ANSI/ASHRAE

[14] ASHRAE is American Society of Heating, Refrigerating and Air-Conditioning Engineers; ANSI is American National Standards Institute.

standard 15 and the European Standard IEC 60335-1 and IEC 60335-2-24. The former are especially important for the operationalization and measurement of values in design, while the second are intended to guarantee that all designs at least meet relevant values to a minimal degree.

Standards often associate certain evaluation criteria and attributes with values like safety (or health or environmental sustainability). They also may contain measurement procedures or criteria for reproducibility (sometimes by reference to other standards). For example, European Standard IEC 60079-20-1:2010 contains the following description of the measurement method for autoignition temperature, an attribute that is relevant for the flammability of a coolant:

> A known volume of the product to be tested is injected into a heated open 200 ml Erlenmeyer flask containing air. The contents of the flask are observed in a darkened room until ignition occurs. The test is repeated with different flask temperatures and different sample volumes. The lowest flask temperature at which ignition occurs is taken to be the auto-ignition temperature of the product in air at atmospheric pressure. (International Electrotechnical Commission 2010a, p. 14)

The same standards also contain the following criteria for reproducibility (and repeatability):

> 7.5.1 **Repeatability**
> Results of repeated tests obtained by the same operator and fixture shall be considered suspect if they differ by more than 2 %.
> 7.5.2 **Reproducibility**
> The averages of results obtained in different laboratories shall be considered suspect if they differ by more than 5 %. (International Electrotechnical Commission 2010a, p. 17)

Even if standards may contain very detailed prescriptions and measurement methods for values, they eventually also rely on value judgments. This becomes quite clear if one looks at the process of standard formulation (and revision). Standards are usually formulated by engineers sitting on standardization committees. The large standardization organizations like ANSI (American National Standards Institute), ISO (International Organization for Standardization), and CEN (European Committee for Standardization) all have procedural safeguards that try to ensure that stakeholders are heard in standard setting and that the resulting standards are based on a certain degree of consensus (or at least a majority). ANSI, for example, has requirements to guarantee openness, transparency, balance of interests, and due process, and standards require consensus. Consensus is defined by ANSI as:

> substantial agreement reached by directly and materially affected interest categories. This signifies the concurrence of more than a simple majority, but not necessarily unanimity. Consensus requires that all views and objections be considered, and that an effort be made toward their resolution. (ASHRAE 2013a)

This is a clear recognition that the process of standard formulation involves value judgments, about which different people (stakeholders) may reasonably disagree. Nevertheless, standardization may be seen as a process in which a certain social consensus is achieved about how to operationalize and measure specific values in the design of specific product classes. If the achievement of this consensus

**Table 2** Properties of refrigerants[a]

| | Environmental sustainability | | Health | Safety |
|---|---|---|---|---|
| | ODP | GWP | Toxicity class | Flammability class |
| CFC 12 | 1 | 10,900 | A | 1 |
| HFC 134a | 0 | 1,430 | A | 1 |
| HFC 152a | 0 | 124 | A | 2 |
| HC 290 (propane) | 0 | 3 | A | 3 |
| HC 600a (isobutane) | 0 | 3 | A | 3 |

[a]OWP and GWP are based on Solomon et al. (2007)

meets certain (procedural) constraints, it might even be the case that a justified consensus is achieved.[15] However, the sheer existence of technical codes and standards should not be seen as the proof that such a justified consensus exists. The coolants' case is an interesting example that shows why such an assumption is problematic.

As we have seen ANSI/ASHRAE standard 15 initially forbade the use of flammable coolants. Table 2 lists a number of alternatives to CFC 12 that were considered. Of these, only HFC134a and HFC152a met the requirements of standard 15; of these, HFC134a quickly became the preferred coolant of the refrigerator industry because HFC152 was moderately flammable. However, the choice for HFC134a was heavily opposed by some environmental groups like Greenpeace, who preferred alternatives with a lower GWP. Eventually this led, at least, in Europe for a choice for other coolants like propane and isobutane.

Interestingly the choice for flammable coolants was accompanied by a change in the relevant codes and standards and in another operationalization of safety. What happened was that safety was at first specified in terms of the evaluation criterion "flammability (of the coolant)." This now came to be replaced by the evaluation criterion "explosion risk (of the refrigerator)." So the prescriptive European Standard EN-IEC 60335-2-24 in its 2010 version now contains the following prescription:

> 22.107 Compression-type appliances with a protected cooling system and which use flammable refrigerants shall be constructed to avoid any fire or explosion hazard, in the event of leakage of the refrigerant from the cooling system.
> Compliance is checked by inspection and by the tests of 22.107.1, 22.107.2 and if necessary, 22.107.3. (International Electrotechnical Commission 2010b, pp. 31–32)

The reformulated standard 15 of ASHRAE in its 2013 version also leaves open the possibility for using flammable coolants, but it does not (yet) provide a new operationalization of safety in terms of explosion risk. Instead it says:

> Group A3 and B3 refrigerants shall not be used except where approved by the AHJ. [AHJ = authority having jurisdiction]. (ASHRAE 2013a, p. 9)

[15]We leave in the middle here when a consensus is justified, but one might think here of John Rawls' idea of an overlapping consensus (Rawls 2001).

If we want to understand why the operationalization of safety in terms of flammability suddenly became contested in the 1990s after it had been taken for granted since at least the 1930s, two contextual factors are of prime importance. One is the growing emphasis on environmental sustainability as an important value (see, e.g., Calm 2008). As we have seen, flammable coolants scored good on this criterion, which raised the question whether a refrigerator with flammable coolant could nevertheless be safe (which is obviously an issue of construct validity). The other has to do with the design of refrigerators. In the 1930s, when the CFCs were introduced, a household refrigerator could contain more than a kilogram of coolant and leakages were not uncommon; at that time explosion risks and toxicity were a serious issue. By the 1990s, after 60 years of design improvement, a typical household refrigerator contained a factor 10–100 less refrigerant, and leakages were much less common. These changes in technical design, in fact, opened the way to another operationalization of safety in terms of explosion risk of the refrigerator rather than flammability of the coolant.

What this story underlines is that changes in context may undermine the construct validity of a certain operationalization and measurement of a value in design. This is largely due to the fact that operationalization of values in design is very context dependent as we have seen. This is different from the case where we measure physical quantities. Of course, it is conceivable that new insights in physics may change the operationalization and measurement of temperature, but the operationalization and measurement of temperature appears to be much more robust against such changes in physics than the operationalization and measurement of values against the occurrence of contextual changes in engineering.

## Conclusion and Discussion

The first conclusion to be drawn from our analysis is that there is a strong analogy between operationalizing physical quantities and moral values in the sense that abstract notions first have to be made more concrete by interpreting them in a specific setting or context: physical quantities in terms of a specific kind of physical system (concrete images) and moral values in terms of evaluation criteria for a specific design. Once that has been done, the interpreted physical quantities and the morally relevant evaluation criteria may be operationalized. In both cases the operationalization thus proceeds in two distinct steps.

But there are also crucial differences. Certain conditions that enable the operationalization of physical concepts in objective measurement procedures are not fulfilled when it comes to the operationalization of values. The most significant difference concerns the embedding of physical and moral concepts in detailed theoretical (abstract) frameworks. Physical concepts are embedded in networks of well-tested theories and operational procedures which make it possible not only to relate various interpretations of a physical concept to each other but also to relate one physical concept to other physical concepts. At present, something similar is lacking with regard to moral values. As a result, issues about construct validity

play no major role in modern physics; the convergence and coherence of theoretical and empirical developments usually makes it possible to settle disagreements about construct validity of a particular physical measurement procedure. Because of the absence of such a network of detailed theoretical frameworks and measurement procedures when dealing with values, issues about whether a particular measurement procedure captures the attribute one intends to measure, that is, issues about construct validity, are much more difficult to resolve. More in particular, we have seen that second-order value judgments play a crucial role in the operationalization of values and that these value judgments seriously undermine any claim that values may be measured in an objective way.

Not only are controversies about construct validity scarce in physics but also controversies about what is called *content validity*. Content validity of a measuring procedure is related to the "adequacy with which the content [of a concept] has been cast in the form of test-items" (Carmines and Zeller 1979, p. 22). Because of the intricate network of theories and measurement procedures in which the notion of temperature is embedded, it is clear that the various notions of temperature employed all hang together and that we are dealing here with a "monolithic" notion that can be measured on *one* temperature scale and that measurement on this temperature scale fully exhausts the content of the notion of temperature. In contrast, the conceptual resources for arguing that a particular specification of a moral value of a design is content valid appear to be missing. If moral goodness of a design is not a monolithic notion, that is, that its meaning is so to speak spanned up by various attributes (dimensions), each attribute corresponding to a different aspect relevant to the moral goodness of a design and being measured on a different scale, then how to ascertain whether all relevant attributes (with their corresponding scales) of the notion of moral goodness have been taken into account? Moreover, how to justify the relative importance of the various attributes for the assessment of the overall notion of moral goodness of a design? In other words, how to aggregate the scores on the various attributes into an overall score for moral goodness? This multi-criteria problem with regard to the morally best design option is just a special case of the general multi-criteria problem that presents itself with regard to selecting the best design option from a set of options given their scores in various criteria (Franssen 2005).[16]

All in all we may conclude that issues about construct and content validity and issues about aggregation in case of multi-attributes make any objective measurement (comparison) of the overall moral value of design options a highly problematic affair. The absence of objective measurement of values, however, does not imply that the operationalization and measurement of values in design is arbitrary. We have seen that technical codes and standards play a major role in the operationalization and measurement of values in design. Although codes and standards ultimately rely on certain value judgments, they may nevertheless establish a reasonable or justified consensus on how to operationalize and measure values in

[16]See the chapter "► Conflicting Values in Design for Values" for a detailed description of this multi-criteria problem for choosing the morally best design option.

design. Standard organizations indeed adhere to certain procedural criteria in order to enable the achievement of such a consensus.

Still, we have seen that we cannot simply assume that current codes and standards establish a reasonable or justified consensus on how to operationalize and measure values in design. One main reason is the highly context-dependent character of operationalizations in design. As a consequence, standards may not reflect the latest technical and social contextual developments, because even if codes and standards are regularly revised, major changes, as with many formalized rule systems like the law, often go slowly and are difficult to achieve. In addition, even if standards can be very detailed and specific for particular kinds of apparatus and devices, they may still not cover all relevant considerations for designing them. For both reasons, the operationalization of values in design processes will usually require value judgments by the designer or design team. However, the value judgments made by designers need not be arbitrary or unjustified. What designers at least can do is to try to embed them in a network of other considerations, including definitions of the values at stake in moral philosophy (or the law), existing codes and standards, earlier design experiences, etc. (For a suggestion on how designers might do so especially when they try to translate moral values in design requirements, see Van de Poel (2014).)

**Acknowledgments** We thank Maarten Franssen for valuable comments on an earlier version of this chapter.

## Cross-References

► Conflicting Values in Design for Values
► Design Methods in Design for Values

## References

ASHRAE (1994) Safety code for mechanical refrigeration, ANSI/ASHRAE standard, 15-1994. ASHRAE, Atlanta

ASHRAE (2013a) Safety standard for refrigeration systems, ANSI/ASHRAE standard, 15-2013. ASHRAE, Atlanta

ASHRAE (2013b) Designation and safety classification of refrigerants, ANSI/ASHRAE standard, 34-2013. ASHRAE, Atlanta

Calm JM (2008) The next generation of refrigerants – historical review, considerations, and outlook. Int J Refrig 31:1123–1133

Carmines EG, Zeller RA (1979) Reliability and validity assessment. Sage, London

Chang H (2004) Inventing temperature: measurement and scientific progress. Oxford University Press, Oxford

Franssen M (2005) Arrow's theorem, multi-criteria decision problems and multi-attribute design problems in engineering design. Res Eng Des 16:42–56

Hunter TA (1997) Designing to codes and standards. In: Dieter GE, Lampman S (eds) ASM handbook. ASM International, Materials Park, pp 66–71

International Electrotechnical Commission (2010a) International standard IEC 60079-20-1 Explosive atmospheres – part 20–1: material characteristics for gas and vapour classification – test methods and data. Edition 1.0 1-2010, Geneva

International Electrotechnical Commission (2010b) International standard IEC 60335-2-24 Household and similar electrical appliances – safety – part 2–24: particular requirements for refrigerating appliances, ice-cream appliances and ice-makers. Edition 7.0 2010-02

Keeney RL (1992) Value-focussed thinking: a path to creative decisionmaking. Harvard University Press, Cambridge, MA

Keeney RL, Gregory RS (2005) Selecting attributes to measure the achievement of objectives. Oper Res 53:1–11

Lelas S (1993) Science as technology. Br J Philos Sci 44:423–442

Nagel T (1979) The limits of objectivity. Brasenose College, Oxford University (May 4, 11 and 18)

Plato, Hamilton E, Cairns H (eds) (1973) The collected dialogues of Plato, vol LXXI, Bollingen series. Princeton University Press, Princeton

Rawls J (2001) Justice as fairness. A restatement. The Belknap Press of Harvard University Press, Cambridge, MA

Sayre-McCord G (2011) Moral realism. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (Summer 2011 Edition). http://plato.stanford.edu/archives/sum2011/entries/moral-realism/

Searle J (1995) The construction of social reality. Penguin, London

Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) (2007) Climate change 2007: the physical science basis: contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge

van de Poel I (2001) Investigating ethical issues in engineering design. Sci Eng Ethics 7:429–446

Van de Poel I (2014) Translating values into design requirements. In: Mitchfelder D, McCarty N, Goldberg DE (eds) Philosophy and engineering: reflections on practice, principles and process. Springer, Dordrecht, pp 253–266

Vincenti WG (1990) What engineers know and how they know it. John Hopkins University Press, Baltimore

# Design Methods in Design for Values

Pieter E. Vermaas, Paul Hekkert, Noëmi Manders-Huits, and Nynke Tromp

## Contents



**Abstract**

In this chapter we demonstrate that contemporary design methodology provides methods for design for moral values. Subsequently, we explore the methodological challenges and problems that this brings to the table. First, we show that

P.E. Vermaas (✉)
Department of Ethics and Philosophy of Technology, Delft University of Technology, Delft, The Netherlands
e-mail: p.e.vermaas@tudelft.nl

P. Hekkert • N. Tromp
Department of Industrial Design, Design Aesthetics, TU Delft, Delft, The Netherlands
e-mail: p.p.m.hekkert@tudelft.nl; n.tromp@tudelft.nl

N. Manders-Huits
Philosophy Department, TU Delft, Delft, The Netherlands
e-mail: n.l.j.l.manders-huits@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_10

contemporary design methods are aimed at realizing values of users and society. These values are in general not moral ones yet do include in specific cases moral values. Second, we introduce a division between *user-driven methods* in which it are the users who introduce the values to be designed for and *designer-driven methods* in which the clients and designers are introducing these values. Third, we discuss two designer-driven design methods in detail for, respectively, design in general and social design in particular: the Vision in Product design method and the Social Implication Design method. Finally, we explore the challenges and problems of design for moral values with these and other design methods. We focus specifically on the designer who, once design is recognized as design for moral values, becomes responsible for the moral values the resulting products have. We argue that in this case the designer should make the moral values of products transparent to clients and users.

**Keywords**
Design methods • Design for moral values • Designer-driven design methods

## Introduction

Analyzing and enabling design for values may be taken as a task that has been born out of the recognition that technical products are not only instrumental means for users to realize goals but also bearers of moral values. This moral ladenness of technology has been demonstrated in philosophy of technology by making explicit the moral values embodied by technical products as diverse as microwaves, tomato harvesters, and obstetric ultrasound scanners (Borgmann 1984; Winner 1980; Verbeek 2011). And once this point was accepted, the next task became analyzing how this morality emerges, to what extent it has its origin in the design processes of the products concerned, and whether these design processes can be adjusted to avoid or steer the moral values that are embodied in technical products.

In this handbook there are ample contributions in which the task of analyzing and enabling design for moral values is taken up within specific technical areas or for specific moral values. In this chapter we take a broader step in understanding this task by considering the possibility of having methods for design for moral values. We demonstrate that such methods are already available in design methodology since contemporary design methods include methods for design for user and social values, which in some cases are moral values. Such contemporary design methods can be divided into user-driven methods in which users introduce the values to be designed for and designer-driven methods in which clients and designers are introducing the values. In this chapter we focus on two designer-driven design methods in detail. Reviewing these design methods as methods for design for moral values, we explore the challenges and problems of design for moral values. By applying these design methods, the designers are introducing moral values in design and become involved into regulating moral dialogues between the clients who commission design projects and the future users who engage with the resulting products.

We start in section "Product Design and its Developments" by introducing current developments in design methodology to focus more on the values users have and on the social values that exist in society. We illustrate these developments by introducing in sections "Design for User Values" and "Design for Social Values" four design methods that support such design for values in different ways. Two of these methods require that designers reason explicitly about values, and we discuss them in detail: the Vision in Product design method (Hekkert and Van Dijk 2011) in section "The Vision in Product Method for Design for Values" and the Social Implication Design method (Tromp and Hekkert 2014) in section "The Social Implication Design Approach to Design for Values." In section "Towards Methods for Design for Moral Values," we give a general model of how moral values are introduced and emerging in design for values. And in sections "Methodological Issues," "Moral Responsibility of Designers," and "Moral Transparency of Designers," we explore methodological and moral issues that arise in design for moral values. Conclusions are drawn in section "Conclusions."

## Product Design and its Developments

Outside their own discipline, product designers are often thought of as either technical engineers or skilled stylists. A designer is either the person who develops the newest tumble dryer at Philips or the one who designs limited-edition Christmas tableware for a department store. This idea that product design is about good functionality and beautiful appearance has shown to be a persistent image of design yet is by now a superseded view. Throughout the last decades, design has developed rapidly from its engineering and architectural roots to a multifaceted discipline. Next to product designers, the world now also knows service designers, co-designers, social designers, eco-designers, and transformation designers, to name a few.

In this development, the design discipline has gradually expanded its focus from the product as such to the human-product interaction, the user experience, and even the implications of this on larger environmental and social systems. Since Norman (1988) introduced the concept of affordance to the design community, the idea that products should be easy to understand and to use has become an important conception in design. Usability studies are executed to understand how people perceive product properties and how these can be designed to guide the actions of the user (e.g., Nielsen 1994). The notion that products should be developed with quite some understanding of the cognitive abilities of human beings therefore is now commonly acknowledged. But not only the user's cognition is important to consider when designing. Emotions too receive increasing attention within product development (Desmet and Hekkert 2007; see also the chapter "▶ Emotions in Design for Values" in this handbook). When being able to trigger specific emotions during user-product interaction, the designer is empowered to induce pleasurable or rich user experiences (Fokkinga and Desmet 2012) or to motivate subjective well-being (Ozkaramanli and Desmet 2012). The common approach in this type of projects is a user-centered one, i.e., an approach in which users are put central in product development processes.

This often means that the designer aims to gain as much insight as possible into the needs, concerns, and values of users and develops a product in line with these.

This increased attention in design for the values of users and of society at large has found its way also to design methodology. In the older engineering design methods (e.g., Pahl et al. 2007), the focus was mostly on functionality. Problems set by clients were interpreted as requests for new functionalities and design processes aimed at developing products that have these functionalities see also the chapter "► Design for Values in Engineering" of this handbook. In contemporary design methods the focus has shifted to also include the values of users. For instance, ethnographic research techniques are incorporated in design methods (e.g., Brown 2009; Plattner et al. 2009) for capturing what values are at play in the problems of users for which is designed and for identifying how those users respond to the products designed for them. And the *Cradle to Cradle* approach to product design (McDonough and Braungart 2002; see also the chapter "► Design for the Value of Sustainability" in this handbook) is a clear attempt towards a design method for the social value of sustainability. Other typical values that are considered in contemporary design are the often-conflicting values of safety and privacy (Friedman et al. 2002). In the next sections we consider in more detail four methods for designing for user values and social values.

## Design for User Values

The idea that design should be sensitive to the values of users has led to an increased emphasis in design methods to understand the concerns of users. One can discern two general approaches towards achieving this understanding and for subsequently addressing them in design. We illustrate these approaches by looking at two design methods: *Participatory Design* and *Vision in Product Design*.

In *Participatory Design* the common idea is that the user is an expert of his own experience and should therefore be incorporated within the design process as such (Sanders and Stappers 2008). The role of the designer then becomes one of facilitator of a process in which people, or expected future users, explain about their (often latent) needs and desires. Several tools and techniques have been developed to support this process (Sleeswijk-Visser 2009). The gathered insights are used as input for the development of product ideas, done by the designer or the design team or done by means of what is called a co-creation process, i.e., a design process in which future users also take part. This process is based on the idea that all people are creative and that it is up to the designers to induce this creativity process with the people for whom they design. This notion is driving the development labeled as "the democratization of design." The designer is no longer the educated and skilled expert designer but acts as the facilitator and coordinator of the design process. The designer no longer develops products that people "consume," but the designer is developing products with the people for whom he or she is designing.

The *Vision in Product* (ViP) design method (Hekkert and Van Dijk 2011) shows a different approach to design for user values. In this method, the designer is

stimulated to understand the future context of the product to be designed. Per definition, people in this future context cannot be interviewed as such. However, the ViP design method does stress the importance of understanding people but, in doing so, emphasizes the importance of the social sciences. The idea is that scientific theories about human beings provide more solid insights of people in future contexts than can ever be gained through interviewing (a few) people in the current context. Another important aspect of the ViP design method is the fact that the designer is asked to explicitly state what he or she wants to offer people in this future context. This statement represents what value the design should have for future users *according to the designer himself or herself*. So although this statement is informed by profound context research, the responsibility for the design is fully placed with the designer. Products are distributed all over the world and can be used by multiple generations, and designers should feel and take the responsibility to foresee the meaning of their designs within those future contexts; designers should not, as is the case in the Participatory Design method, distribute this responsibility to a few people who consider this meaning in a present context.

## Design for Social Values

In addition to design for user values, an interest to also design for social aims has rapidly increased in the last decade. Designers seem to have a growing motivation to not solely develop consumer products but to employ their talents and skills to "do good." One of the ideas is that design can improve the lives of many who are living in developing countries. This is, for instance, done by developing products that empower people in developing countries (e.g., by designing for the Base of the Pyramid (BoP), Whitney and Kelkar 2004; or by applying a so-called capability approach, Sen 1999; Nussbaum 2011; see also the chapters "▶ Design for the Value of Sustainability" and "▶ Human Capabilities in Design for Values" in this handbook). But also in Western societies, design has received attention for its potential to induce social change. For instance, the London-based agency Participle involves users in the development of products to address social issues.[1] CEO and president of design consultancy IDEO Tim Brown (2009), for instance, advocates the process of design as a tool that can help organizations and societies change. This development towards design for social aims can be seen as one towards design for social values, being values that groups of individuals or societies have and that express what they hold as good for society. Examples of such social values in design are cohesion, equality, safety, sustainability, and participation or inclusion. Two approaches to what can be called "social design" are *Transformation Design* and *Social Implication Design*, and they can be seen as extensions of, respectively, the Participatory and Vision in Product design methods.

[1] http://www.participle.net/. Retrieved 12 November 2012.

*Transformation Design* (Burns et al. 2006), or what some explain as service design applied to social systems (Saco and Goncalves 2008), is driven by the notion that design skills and techniques can be extremely valuable in changing social and public services (Sangiorgi 2011). Because public services are usually developed through a top-down approach and driven by political considerations, these services often fail to answer users' needs and concerns. The user-centered focus of designers and the skill to translate user concerns to concrete solutions therefore logically open up new opportunities to improve services in health care, education, and politics. In order to realize insights in user needs and desires, the approach of Transformation Design shows similarities with the Participatory Design method. Transformation Design also emphasizes the role of the users in the design process in order to better answer their needs and desires. Moreover, when applied to induce social change (e.g., Manzini and Rizzo 2011), the idea is that by distributing the power to users within the decision-making processes, the capabilities of these users grow, and the process of change and development will proceed as an ongoing process.

*Social Implication Design* considers each design as a means to change and shape behavior and thereby induce social change (Tromp et al. 2011; Tromp and Hekkert 2014). Social issues are often mentioned and discussed in terms of problematic behavior that needs to be changed, e.g., health issues refer to people eating unhealthily or not exercising, crime issues point at criminal activities that need to be stopped, and immigration issues refer to lacks in integration efforts by immigrants. The Social Implication Design approach supports designers to analyze these relations between social issues and human behavior and use these relations in their designs. Based on what is desired from a social perspective, the designer is asked to shift to a user perspective in considering how this social change can be best achieved by stimulating different behaviors and how this stimulation can be best achieved by means of a design. Social Implication Design is based on the ViP design method and therefore stresses the responsibility of the designer in questioning, discussing, and finally defining what is a desired social change and in questioning and analyzing the generally assumed relations between social issues and behavior of people. By taking this responsibility and evoking their creativity, designers are to create both well thought-through and effective designs in realizing change. Social Implication Design rather stimulates designers to create the optimal conditions to induce sustainable social changes in the longer term, instead of focusing on inducing change with people in the current context (Tromp and Hekkert 2010).

The four described approaches to design can be ordered using two divisions (see Fig. 1). First, Participatory Design and Vision in Product Design are concerned with creating products and services for users and thereby take a user perspective, whereas Transformation Design and Social Implication Design are focused on realizing societal aims and therefore take a social perspective to design. Second, these approaches define the role of users differently. Participatory Design and Transformation Design both are user driven by taking users as the experts of their needs and goals and of their experiences with designs. In contrast, Vision in Product design and Social Implication Design both are designer driven by considering designers as having the final responsibility for determining the values designed

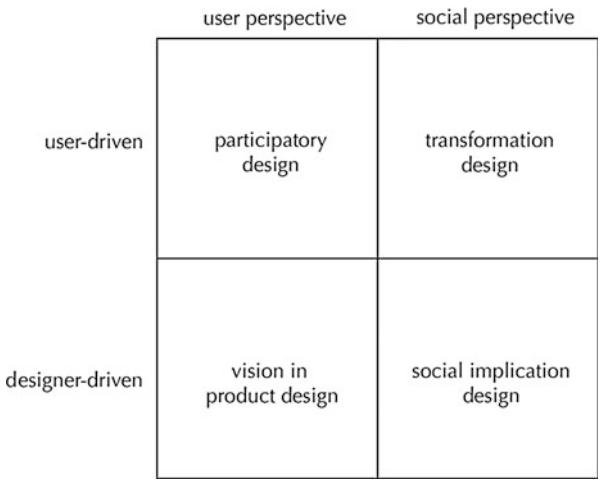



**Fig. 1** Four methods to design for values

products have and the experiences these products will induce with users. In these designer-driven approaches, it is the designer's task to convince future users of the meaning or values embedded in their design; the importance that designers understand human beings is emphasized, and for this understanding these approaches rely more on the social sciences than the user-driven approaches. Especially when designers are concerned with the long-term consequences for society, like in the Social Implication Design approach, the designer cannot afford to rely on user responses to a current context. In the following two sections we continue with describing how designers bring in values in their designs in the two designer-driven approaches; further descriptions of the user-driven approaches can be found in the chapter "► Participatory Design and Design for Values" of this handbook.

## The Vision in Product Design Method for Design for Values

How then do designers deal with values and design for values with current methods? For taking up this question we describe in some detail how the design process is structured by the Vision in Product and Social Implication Design methods; both these methods make designers explicitly think about and formulate the values they incorporate in products. In this section we describe a case of design for values with the ViP design method, and in the next section we consider the Social Implication Design approach, again by using a case.

In 2009, Anna Noyons graduated at the Faculty of Industrial Design Engineering of Delft University of Technology on a project for HEMA, a popular, low-end Dutch department store for basic home products, such as stationary, kitchenware, care products, and food. Surfing on the trend of sustainability with their new brand "naturally HEMA," Anna's assignment was to design a new sustainable product

(line) with a focus on bio-based plastics. In this way, HEMA aimed at positioning itself as a company that cares for the environment and our ecosystem.

In order to avoid "simply" turning an existing product line into a sustainable, bio-based version, Anna decided to apply the Vision in Product (ViP) design method (Hekkert and Van Dijk 2011) to force her to take a step back and define a holistic vision on a sustainable future. This ViP design method explicitly forces the designer to take such a holistic perspective by first building a future *context* to which the final design must form an appropriate response. This context is built around a set of insights and observations – called "factors" – that together define how the designer sees the world of the domain chosen. Whereas the selection of factors is (partly) ruled by the values the designer holds, the factors themselves are mere observations and do not include value judgments. This judgment is postponed to a later stage at which the designer defines a statement that covers his or her position towards the context. This statement has the form of "what do I as designer want to offer people given this world?," and the propositional nature of this statement expresses three core values that in the ViP design method are claimed for the designer: the designer needs *freedom* to act, to act *responsibly*, and to make an appeal to his true, *authentic* self.[2]

In order to develop her context, Anna took a second look at her brief. The initial brief of the client is often too narrowly defined and based on a set of constraints that may appear invalid on second thought. In her case, the brief was too open and needed refinement. To restate the brief, Anna analyzed the concept of sustainability and carefully examined the HEMA brand. Instead of (only) seeing sustainability in terms of product qualities, such as the use of bio-based plastics, Anna adopted the more holistic approach from Ehrenfeld (2008). According to Ehrenfeld, sustainability is the possibility for human and other lives to flourish on this planet forever; we need to think global and act local. This local focus guided Anna from sustainability to a second value, which she took as a starting point for her context research: HEMA is at the center of local community life and could foster local communities and social cohesion. Finally, Anna argued that HEMA should depend less on base materials and gradually focus itself more on service design, reducing waste, and environmental costs in production.

The initial findings and observations of Anna concerning the brief were put aside as constraints. Whenever appropriate, these constraints could affect various subsequent decisions and thus reenter the design process. Values are often implicit in a design assignment, and the ViP design method forces you to make these explicit and/or redefine them. Anna's interpretation of the brief reveals her "truth," her vision on what is good or bad for the company given the direction (of sustainable design) chosen.

[2]These three values of freedom, responsibility, and authenticity are values for the designer; by embracing these values for the designer, the products designed with the ViP method do not necessarily also embody these values.

Both from a strategic, company perspective and a user perspective, Anna chose to refine the brief and focus on baby-hardware products, keeping the sustainability goals in mind. More specifically, she decided to define her domain as "the maternity period." For this domain she selected factors she considered relevant and personally interesting, among others, by interviewing young mothers and exploring their world, often perceived as "narrow." Some of the factors she selected were:

- (Pregnant) women expect (and are expected) to be the happiest person in the world, feeling guilt when experiencing doubt or other negative emotions.
- Within society and relationships, men and women are becoming more equal.
- People are becoming more aware of the effects of industrialization.
- Pregnancy = immobility.
- Fathers feel left out during maternity.

Note that some of these factors concern principles, more or less fixed patterns of nature or psychology; others refer to things that are changing, such as trends or developments. These two types typically constitute the building blocks of any context. As said, these factors do not indicate what (according to the designer) should happen, but reveal how he/she looks at the world around the domain.

After all factors have been brought together into a single, unified view of the world, a process that requires time and a range of design skills, the designer can start to formulate how he/she wants to respond to this world. In Anna's project this statement was "I want (future) parents to feel encouraged to surrender to the smaller world associated with the maternity period, to build a trustworthy base together with, and for their child to be raised in." Clearly, the designer must consider his or her own value system to come to such a position; the ViP design method deliberately forces designers to take a stand and play out their value systems. The dominant (and related) values/beliefs of Anna that made her take this particular position were:

1. People should live in the here and now and cherish the moment.
2. One must submit to the situation and accept that things are as they are.

This statement is the first part of the vision and sets the goal for the final design.

Products get their meaning in interaction. In order to see how design can fulfill the goal stated, the ViP design method demands the designer to first conceptualize this interaction. These interaction qualities also carry implicit values, but these are goal-directed values. In Anna's case, two of these were "conscious devotedness," as present in knitting a vest, and "modest," the interaction should not ask too much, just enough. The important thing to see here is how these interaction qualities – as abstract as they are – allow you to surrender to a smaller world. Subsequently, after the interaction has been defined that can realize the goal, product qualities can be defined that should bring about the interaction. Anna argued that her "product" (note that at this stage it could still be any solution), for instance, should "invite to act consciously" and "be naturally safe."


**Fig. 2** The final product concept

With this vision in mind, Anna designed a product-service system consisting of various building blocks (e.g., biodegradable containers for paint, yoga pillow; see Fig. 2) and a website to share experiences with women in the same situation in their neighborhood. Again, some personal values can be identified that brought Anna to this solution:

1. Products (should) have a story: products should not be overly defined, but people need to find out themselves what they mean and how they could be used.
2. Product longevity: things must have a permanent, long-lasting value.

Summarizing, the ViP design method not only allows designers to demonstrate values in the process, it actually demands them to do so. Designing is (always) about making choices and setting a future belief that people (users) must eventually embrace. This simply cannot be done without including certain values and abhorring others. This insight resulted in an approach that requires the designer to explicitly define and execute (personal) values, reason about these values, turn them into design decisions, and maybe most importantly, take responsibility for the result and what it brings to people.

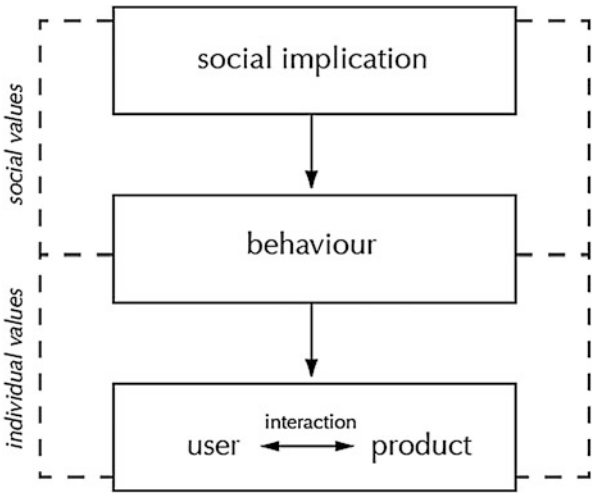



**Fig. 3** The role of values in the Social Implication Design approach

## The Social Implication Design Approach to Design for Values

In 2011, Sacha Carina van Ginhoven graduated at the Faculty of Industrial Design Engineering of Delft University of Technology on a project for Stichting Boog, a foundation that develops social projects in neighborhoods in the city of The Hague. In these neighborhoods youth is dominating the streets by hanging around and committing crimes such as burglary and vandalism. Stichting Boog is concerned about the tensions that this behavior induces in the neighborhood but experiences difficulties in getting into contact with these youngsters, let alone, realizing any change in their behavior. A design student was invited to examine the situation and develop a product or service which would improve the situation.

In her project, Sacha applied the Social Implication Design approach (Tromp and Hekkert 2014). This approach supports designers to develop products and services based on their intended implications for society. In addition to explicitly stating what the designer wants to offer the future users of the design, like in the ViP design method , the designer needs to explicate what social values he or she considers important. Given the fact that products and services influence people's behavior and thereby cause implications for society, the Social Implication Design approach supports the designer in reversing this causality. To induce an intended, *desired* implication for society, what behavior needs to be stimulated and by means of what product can this best be achieved? In the Social Implication Design approach, the designer is asked to study social values to understand which implications are desirable and to study individual user values to understand through what type of product this can best be achieved (see Fig. 3).

To understand and explore the situation at hand, Sacha spoke to youngsters in the streets, made observations with police officers, and carried out literature research.

She found that the youth resemble what is called a "street culture," representing a set of norms and values and fashion styles and using typical "slang." Youth often feel distrusted and rejected by society, and their street culture is an expression of their nonconformism to the standards of society. Within a street culture like in The Hague, having money or expensive clothing and goods is highly valued. These youngsters, often raised in poor families, care a lot about expressing their wealth. As a consequence, this makes them vulnerable to the temptation of criminal activities. Yet in the interviews, Sacha found that when she asked them about their desired future, they all mentioned very conventional aspirations like having a family, a house, and a nice job. The paradox of how society reacts to these youngsters is that, on the one hand, governments put in a lot of effort to keep them "on board," preventing them to slide into criminal circuits, while, on the other hand, many people have prejudices against these youngsters, and employers often feel hesitant to hire them.

From a social perspective, Sacha stated that she wanted to integrate the street culture better in our society, stressing inclusion, and she regarded work mediation as the way to do so. From a social perspective "work participation" of youngsters is desired in both a social and an economic sense. It decreases the chance that youngsters will employ criminal activities. From the perspective of the youngsters, work leads to money to spend and therefore offers a way to gain individual status and respect. But this only holds when the job as such does not exclude them from their peers in the streets. Currently, the reason not to apply for a job is the chance of getting rejected both by society and by their peers. Therefore, Sacha did not want to stimulate youngsters to adjust to society, but wanted society to move a bit closer to the streets.

The ultimate behavior Sacha aimed to influence is the youngsters' criminal activities. However, instead of discouraging youngsters by pointing to the downsides of engaging in criminal activities, she set out to stimulate desired behavior by offering competitive benefits for youngsters to the ones that arise from criminal activities. She decided that the activity of applying for jobs was the most fruitful to stimulate. To understand how to do so, Sacha disentangled the job application process and the concerns of youngsters within this process. An important part of a successful job application is making a good first impression. However, youngsters often experience difficulties in communicating the right image. On the one hand, youngsters often lack the knowledge of social norms in application procedures. On the other hand, recruiters may be negatively biased in judging these youngsters. Youngsters therefore experience fear to apply for a job. To overcome this, Sacha wanted to enable youngsters to give this first impression in a way that they would feel comfortable with and which would reveal an honest reflection of their identity to recruiters.

The WorkTag is a sticker with a QR code. The idea is that employers can apply this tag to the place where work is available, e.g., near a bus stop when there is a vacancy for a bus driver, in a park to recruit gardeners, or near a construction site when there is a need for construction workers (see Fig. 4). When screening the tag with their smart phones, more information about the job is given. Not only can these tags lead to permanent work but also to instant and short-term chores, like helping


**Fig. 4** The WorkTag ("*Ik heb werk voor je*" means "I have a job for you"; "*Ik heb een stage voor*" means "I have an apprenticeship for you"; "*Ik heb 50 euro voor je*" means "I have 50 euros for you")

somebody in the neighborhood with painting their window frames. When youngsters are interested in the job, they are invited to record a small video to express their interest. The application supports them in doing so. The recruiter invites job applicants on the basis of this video. The idea is that by integrating these videos into the job application process, recruiters receive a more honest image of the youngsters. On the basis of these videos, recruiters select whom to invite for a meeting.

Bringing a job application's first impression literally to the streets and by providing youngsters the means to apply for a job in a way that suits them, Sacha aimed to seduce youngsters to start working and society to accept youngsters as they are.

Sacha's project approach forced her to question the conventional ideas about criminal youth. She critically examined the relationship between youngsters in the street and society and questioned the assumption that youngsters, i.e., the "misfits," need to adjust to societal standards. In contrast, she aimed to challenge society to judge youngsters differently. Although this statement is a personal statement and therefore evidently includes personal values, she examined the values we have as society and agreed with the fact that "participation" is desired from a social perspective. However, in her attempt to lead youngsters to work, she related as much as possible to the concerns and values of youngsters. The Social Implication

Design approach explicitly requires a dual perspective from the designer: a social perspective and a user perspective. In taking these two perspectives, the designer is asked to consider both social and user values and to explicitly reason about them. However, it is up to the designer how to address these values. It is by taking this position that personal values unavoidably enter the design process. Simply by openly stating one's position, this position can be questioned and discussed and thereby increases responsibility with the designer.

## Towards Methods for Design for Moral Values

The examples of the Vision in Product and Social Implication Design methods show that designers have methods that support them in design for user and social values. These methods contain stages in which designers explicitly identify the values inherent to design assignments or formulate the values by which they plan to address the assignments. These values are in general user and social values but can in specific cases also be moral values. For instance, the design cases discussed above concern the moral values of sustainability and of participation. And the user and social values for which designers design include the moral values of safety, privacy, and equality. Hence, the Vision in Product and Social Implication Design methods can be used to design for at least some moral values.

As this shows that contemporary design methods can support design for moral values, we can take up our next task of exploring more generally what methodological problems and challenge design for moral values possess. In this exploration we focus on the role of the designers in design for moral values, who become actors who can actively introduce moral values in the design of products, who become responsible for the moral values the resulting products have, and who, as we argue, should make these values transparent. For this exploration we use a simple model of how moral values are introduced and emerging in design for values, as depicted in Fig. 5, and again draw from experiences with the ViP and Social Implication Design methods.

Let $V_C$ be the moral values that are specified by the client in the initial design description. Let $V_P$ be the moral values that are added during the design process.

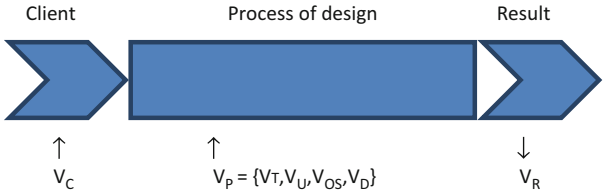



**Fig. 5** A simple model of how moral values are introduced and emerging in design for values

Let $V_R$ be the moral values that the designed product eventually holds. The moral values $V_P$ added during the design process may be divided in four categories:

- Moral values $V_T$ that are required by technical regulation and legislation (say, safety and, increasingly, sustainability).
- Moral values $V_U$ that are brought in by users through their active role in the design process or through user feedback as organized by the designer.
- Moral values $V_{OS}$ that are brought in by or for other stakeholders in the design process, as people who are not using or ordering the design product but are affected by their existence or operation.
- Moral values $V_D$ that the designer brings in personally.

(It may be argued that the last set $V_D$ includes much of the first three; most moral values are brought in in the design process by the designer, as it is the designer who decides that specific technical values $V_T$ apply or that specific values $V_U$ and $V_{OS}$ of users or other stakeholders are to be included or not. This argument makes sense and gives rise to questions about the responsibility of designers to regulate the inclusion of moral values in design (see section "Moral Responsibility of Designers"), yet in the model $V_D$ is meant to refer to those moral values the designer brings in on the basis of personal considerations or preferences.)

One can now discern three types of issues related to design for moral values: methodological issues about design for moral values, issues about the moral responsibility designers acquire as initiators and regulators of moral values in design, and issues about the transparency of design for moral values.

## Methodological Issues

A first methodological issue that can be raised for design for moral values is that it assumes that designers have all kinds of new design abilities. Regular engineering design, when taken as design for functions and structural properties, is supported by a broad set of tools for creating products with specific functions and structural properties and by clear criteria for determining and arguing that the resulting products have these functions and properties (e.g., Pahl et al. 2007). The idea of design for moral values extrapolates this engineering model of design to the moral domain, suggesting that designers have also the tools and criteria for letting products have specific moral values: in terms of the model depicted in Fig. 5, designers are assumed to be able to deliver products that have as their moral values $V_R$ precisely the values $\{V_C, V_T, V_U, V_{OS}, V_D\}$. One can however question whether designers have these tools and criteria. The Cradle to Cradle design method (McDonough and Braungart 2002) may be taken as one that spells out tools and criteria for designing for the value of sustainability (see also the chapter "▶ Design for the Value of Sustainability" in this handbook). But contemporary design methods do not give such means for all possible moral values; methods such as ViP and Social Implication Design provide merely general guidelines for when and

how to identify and reason about (moral) values in design. Criteria for determining whether products have specific moral values are, moreover, not up front fixed. Moral values are typically initially formulated in generic terms, and designers should give meaning to them and embody them, as is illustrated by the two design cases that were given above. Designers should explore what users consider important moral values, as in the Participatory and Transformation Design methods, or designers should analyze future contexts for the products and then specify what moral values are at stake, possible with the help of the social sciences, as in ViP and Social Implication Design methods. The question how to design for specific moral values is taken up in other contributions to this handbook. The operationalization of moral values is discussed in the chapter "▶ Design for Values and the Definition, Specification, and Operationalization of Values."

A second methodological issue is that the values $V_C$, $V_T$, $V_U$, $V_{OS}$, and $V_D$ for which is designed may draw in opposite directions, leading to the moral problem of reconciling conflicting values or of finding acceptable trade-offs between them. A simple example may be the over-dimensioning of products and components. This over-dimensioning may be a design solution to meet the values of robustness, safety, and reliability but can easily be in conflict with the technical value of efficiency or the moral value of sustainability. The issue of conflicting values is often present when designing for social problems, as social problems arise when many people act in ways that may benefit themselves (contributing to personal values) but which causes drawbacks for society (conflicting with social values). For instance, taking the car to work is something that contributes to personal values of freedom, independence, and comfort but conflicts with social values like sustainability. Interestingly though, design may offer unique solutions to such conflicting concerns. Just like the designer's integrative thinking is an important and unique skill in overcoming clashes between, for instance, the aesthetics and ergonomics of a design (Dorst 2007), so can they employ this skill to realize designs to overcome these conflicting concerns (Tromp and Hekkert 2010). Conflicting concerns in design for values is discussed in more detail in the "▶ Conflicting values in Design for Values" chapter.

Effective design for moral values means, finally, that the values $V_R$ a product eventually has are exactly the values for which it was designed, which implies that the product should not embody moral values that are not included in $\{V_C, V_T, V_U, V_{OS}, V_D\}$. In the description of the responsibility of the designer as given by the ViP design method, section "Design for User Values," it was already noted that this effectiveness requires that future contexts of products are well thought through by designers. Yet, a product can still start to embody new moral values when it is in use. In analyses in philosophy of technology, it is, for instance, argued that products sometimes act as moral agents themselves and by mediation pick up new moral meanings (Verbeek 2011). In principle this adoption of additional and unintended moral values is not a new phenomenon in design, since also on the functional level products may acquire unintended new functions in their use. Yet, by this phenomenon design for moral values may be less effective than expected, defining a third methodological issue. This issue is analyzed in more detail in this handbook in the chapter "▶ Mediation in Design for Values."

## Moral Responsibility of Designers

When design methods for moral values become increasingly available, and designers do acquire the tools to effectively let products have moral values, then the role of designers changes substantially. Designers then not only can solve practical problems for clients, users, or society by creating products with specific functions and structural properties but also address moral issues by creating products that embody moral values. Moreover, designers become agents that help clients, users, and society to include their moral values in products, and designers become agents that may themselves include moral values in these products. These new abilities will lead to additional responsibilities of designers and put engineering ethics in a new perspective. Designers are then not only responsible for possible physical consequences of the use and misuse of the products they create but become also directly responsible for the moral values products have. The instrumental view that designers merely create products and that clients and users are morally responsible for the way they employ these products becomes with design for moral values even more untenable as it already is (Vermaas et al. 2011), for now designers do know what moral values they design for and not design for. Guns, to return to the classical example, become with the possibility of design for moral values, products explicitly designed for the (moral) values of public safety and policing, or may be designed for the (moral) values of self-defense or even random attack. This responsibility of designers for the moral values products have has different forms. First, there is a responsibility for the moral values designed for. In terms of the values $V_R = \{V_C, V_T, V_U, V_{OS}, V_D\}$ discerned in the model, one may assume that design for technical values $V_T$ like safety and sustainability does not raise moral questions. But design for client values $V_C$ and user values $V_U$ raises moral issues about what values designers can design for, as clients and users may also have morally bad values they want products for. Second, designers become responsible for the inclusion or exclusion of the moral values $V_C$, $V_U$, and $V_{OS}$ of clients, users, and other stakeholders. Third, designers are morally responsible for the values $V_D$ they introduce themselves. And finally, when products pick up unintended moral values other than those in $\{V_C, V_T, V_U, V_{OS}, V_D\}$ for which they are designed, the designers may be held responsible for delivering morally faulty products. Hence, with the rise of methods for moral values and the moral responsibility it brings, it may be argued that also new professional codes of conduct and additional legislation are called for that regulate these new responsibilities.

That designers themselves can introduce moral values $V_D$ in design was shown by the cases described in the previous sections. In the case described in section "The Vision in Product Method for Design for Values," the client HEMA had defined "sustainability" as the leading value, yet it was the designer who chose a particular reading of this moral value and added the value of local social cohesion. In the Stichting Boog case, described in section "The Social Implication Design Approach to Design for Values," the assignment of the client was merely to change specific behavior, and it was the designer who introduced the moral value of participation. Finally, in design for social values, as discussed in

section "Design for Social Values," designers explicitly chose themselves societal values to design for because they want to "do good." This introduction of moral values $V_D$ in design is related to the designer-driven nature of the ViP and Social Implication Design methods used in these cases yet may become a standard part of design methods for moral values. In current design methodology the role of designers is already acknowledged to be an active one. All methods promote designers to co-determine the requirements a product has to meet: design then starts with a client who gives a first description of the product to be designed, after which the designer helps to further specify this description by analyzing the client's goals and descriptions. The designer knows what is technologically feasible, what is required by regulations and law, and what makes the product usable and safe. Moreover, the designer may add technical or aesthetic elements. Finally, the designer may even propose to change the initial description of the client. If a set of requirements derived from the initial description of the client defines a design task which proves impossible to carry out or which leads to unattractive solutions, the designer may propose to reframe the design task (Schön 1983; Cross 2006). This implies that the client's goals are attempted to be realized by products that are different from the ones initially described by the client. In product design this active role of the designer to co-determine or reframe the requirements of the products to be designed is considered to be a positive one: it leads to efficient, user-friendly, and safe products and, in the case of reframing, to innovative products. In design for moral values, designers may be expected to also take this methodological role of actively co-determining or reframing the moral values that products are supposed to have to arrive at innovative solutions, as was illustrated by the Stichting Boog case. Again this possibility of design for moral values leads to moral issues about what values designers may introduce, for also these values $V_D$ may be morally wrong. This possibility also leads to moral issues among designers and clients, users, and other stakeholders, which is a final topic we explore.

## Moral Transparency of Designers

When designers actively introduce moral values $V_D$ in their designs, moral conflicts with the clients, users, and other stakeholders may emerge. For exploring such conflicts one can discern two separate reasons designers may have for introducing moral values $V_D$ in their designs. These two reasons are not exhaustive

The first reason for a designer to personally introduce moral values may be that they are *goals* of the designer himself. A designer may be carrying out a design task set by a client and add moral values to the product to be designed which are neither wished for by the client nor instrumental to the realization of the client's wishes. A case that would illustrate this first possibility is one that is well known in philosophy of technology, namely, the case of the Long Island overpasses. In the design of the Long Island's parkway system, the designer Robert Moses chose allegedly low overhang bridges for racist reasons; the low overhang bridges would effectively obstruct poorer minorities from traveling through the parkway system,

assuming that they travel by public buses (Winner 1980). This would mean that Moses added the value of racism to his design.[3]

From a moral point of view, one could argue that designers should not bring in morally bad values like racism. But even if these values would be morally acceptable, one could argue that designers should not unconditionally bring in moral values for their own goals. In this case, designers breach the trust clients put in designers and are violating the moral autonomy of the client and of the users of the resulting products. Yet, in design methodology designers have an *active* role of co-determining or reframing the requirements a product has to meet, and by the design methods discussed in this chapter, designers are actually stimulated to bring in their own moral values. A more subtle response would be to require that when designers do bring in their own moral values, it should have the nature of a proposition to clients, users, and other stakeholders, allowing their acceptance of or informed consent with regard to the introduced values $V_D$. Informed consent means in this context that people are informed that the designer of a product added moral values and that people are put in the situation that they can make up their minds about whether they accept these products with these values (assuming that people can chose alternative products). In fact, in the discussed ViP and Social Implication Design methods, designers are actually stimulated to explicitly and clearly argue for the moral values they introduce.

The second reason a designer may have for introducing moral values $V_D$ in design is to let them serve as *means* for realizing the purposes of the client. A designer may be faced with a specific design task by the client, which may or may not contain other values $V_C$, $V_T$, $V_{OS}$, and $V_U$, and decide to find a solution to it by designing a product that has also moral values $V_D$. The Stichting Boog case of section "The Social Implication Design Approach to Design for Values" may serve as an example. The client wanted a product to change the behavior of youngsters, and the designer introduced the moral value of participation to realize this goal. Using moral values as a means in design may be innovative in design and broaden the spectrum of tools designers have for realizing the goals of their clients. Yet this second possibility again raises the issue of whether the clients, users, and other stakeholders could agree with these values. In the Stichting Boog case, the use of moral values as a means seems morally acceptable towards the client, since the value of participation may have been one that is held by the client. Yet, when considering the youngsters involved, one can raise worries against carrying out the design for moral values in this way. If these youngsters start using the service without being aware of the inherent goal of realizing their participation, then ethically speaking their autonomy is violated; no informed consent has been given by these youngsters, ignoring the possibility that they may not endorse participation. Hence, when designers introduce values $V_D$ in design, even if merely as a means, then morally speaking this requires transparency regarding the inclusion of these values to all parties concerned.

[3]This account of the design of the Long Island overpasses is contested (Joerges 1999).

One could argue that this requirement of moral transparency will become in part incorporated in design methods for moral values. Given current design methodology this transparency is easily achieved towards clients. In current design methods the interactions between the client and designer are already one of informed consent due to a constant conversation between the two: clients present their initial design assignments to designers, and designers discuss with clients the ways in which they develop these assignments and eventually resolved them in design. Design methods for moral values may therefore be expected to include ample interaction with the client to also reach agreement about the moral values introduced by the designer during the design process. Still, additional work has to be done to add this interaction between designers and clients to design methods for moral values. Conceptualizations of moral values are often essentially contested. These need not be thorough philosophical disputes between the clients and designers, but rather minor differences of opinion which may nevertheless gradually influence design choices (including different ways of taking a certain product into use). Designers are now required to guide the process of capturing moral values in concrete design decisions.

Moral transparency is harder to achieve towards the users. In section "Design for Social Values" we made a distinction between user-driven and designer-driven design methods. The Participatory and Transformation design methods were taken as user driven since by these methods designers become facilitators and coordinators of the design process. The ViP and Social Implication Design methods were designer driven since in these methods designers are taking responsibility to take decisions for users. One could argue that in user-driven methods, in contrast to designer-driven methods, any moral value introduced by the facilitator designer is introduced in a transparent manner, since it are the users participating in the design who eventually adopt or reject the introduced values. Yet, this argument fails since the users participating in the design process are just a few and cannot represent *all* the prospective users of the designed product. Hence, even if the users participating in a design process give their informed consent, the future users have not yet done so. This limitation that designer can interact only with a few users is in fact the motivation behind designer-driven methods to lay the responsibility for introducing values in designs with the designers involved. Still, also for these methods the conclusion is that designers then introduce moral values without having informed consent by *all* possible and future users of the designed products. Moral transparency of design for moral values towards users can then at most be achieved by requiring that designers deliver their products with a clear specification of the moral values they are designed for, such that future users have at least the possibility to inform themselves about these values.

Consider, as a final example, the design of a baby stroller – the *Bugaboo* – developed by the professional designer Max Barenbrug. At the time of its inception, baby strollers were mainly foldable buggies under the assumption that parents wanted baby strollers cheap, instrumental, and easy in use. Analyses similar to those that are part of the ViP design method revealed that there was a new group of parents which had little time to spend with their children – e.g., couples where both

partners have jobs – and who felt guilty for this. Partly due to these guilt feelings, these customers were ready to spend more money on products for their children and were susceptible for strollers providing comfortable and protective environments for their babies. The result of the design by Barenbrug was a stroller that was more luxurious and five times more expensive than regular buggies (Hekkert and Van Dijk 2011, pp. 136–137). When seen as a product designed for the values of comfort and protection, the Bugaboo may be taken as one in which transparency with respect to the embodied (moral) values has been achieved, since these values were communicated and the identified group of users shared them. But when seen as a stroller designed for the values of extensive parental care and of relieving the guilt of parents with double jobs, a different perspective emerges. The identified group of parents may also have had these values, but it can be doubted whether these values were also made explicit to the parents and whether these parents would accept the stroller as embodying them: the Bugaboo did not buy these parents more time with their children, as, say, washing machines did.

## Conclusions

Current design methods can already be taken as enabling design for moral values, and although they cannot be taken as full-fledged design methods for moral values, they show that such methods exist and may become increasingly available.

We presented two design methods in detail – the Vision in Product and Social Implication Design methods – in which designers reason explicitly about moral values and in which they introduce moral values themselves. Drawing from these two example methods, we explored the methodological and moral issues that emerge when arriving at full-fledged design methods for moral values. The methodological issues are the development of tools for design for moral values, the operationalization of moral values, the resolution of conflicts between moral values, and the emergence of unintended moral values the designed products may acquire in future use. The moral issues included the responsibilities designers acquire for regulating the moral values clients, users, and other stakeholders propose for the products to be designed and the moral values designers may introduce themselves in designing. Current design methods already presuppose and stimulate that designers actively co-determine or reframe the design assignments they take up, and design methods for moral values will equally give designers such an active role in moral designing, as was illustrated by our examples.

Finally it was argued that designers should be transparent about the moral values they introduce in design and should aim at informed consent towards clients and users. Design methods for moral values will have to support designers in conceptualizing moral values in their interaction with clients and users partaking in the design processes and to provide the tools to communicate the moral values products have to future users not partaking in the design processes.

## Cross-References

- ▶ Conflicting Values in Design for Values
- ▶ Design for the Value of Sustainability
- ▶ Design for Values and the Definition, Specification, and Operationalization of Values
- ▶ Design for Values in Engineering
- ▶ Emotions in Design for Values
- ▶ Human Capabilities in Design for Values
- ▶ Mediation in Design for Values
- ▶ Participatory Design and Design for Values

## References

Borgmann A (1984) Technology and the character of contemporary life: a philosophical inquiry. University of Chicago Press, Chicago
Brown T (2009) Change by design: how design thinking transforms organizations and inspires innovation. Harper Business, New York
Burns C, Cottam H, Vanstone C, Winhall J (2006) Transformation design. RED Paper Design Council, London
Cross N (2006) Designerly ways of knowing. Springer, London
Desmet PMA, Hekkert P (2007) Framework of product experience. Int J Des 1(1):57–66
Dorst K (2007) Design problems and design paradoxes. Des Issues 22(3):4–17
Ehrenfeld JR (2008) Sustainability by design: a subversive strategy for transforming our consumer culture. Yale University Press, New Haven
Fokkinga SF, Desmet PMA (2012) Darker shades of Joy: the role of negative emotion in rich product experiences. Des Issues 28(4):42–56
Friedman B, Kahn PH, Borning A (2002) Value sensitive design: theory and methods. University of Washington technical report, 02-12
Hekkert P, Van Dijk M (2011) Vision in design: a guidebook for innovators. BIS Publishers, Amsterdam
Joerges B (1999) Do politics have artefacts? Soc Stud Sci 29:411–431
Manzini E, Rizzo F (2011) Small projects/large changes: participatory design as an open participated process. CoDes Int J CoCreat Dese Arts 7(3–4):199–215
McDonough W, Braungart M (2002) Cradle to cradle: remaking the way we make things. North Point Press, New York
Nielsen J (1994) Usability engineering. Morgan Kaufmann, San Francisco
Norman DA (1988) The psychology of everyday things. Basic Books, New York
Nussbaum MC (2011) Creating capabilities: the human development approach. The Belknap Press of Harvard University Press, Cambridge
Ozkaramanli D, Desmet PMA (2012) I knew i shouldn't, yet i did It again! emotion-driven design as a means to motivate subjective well-being. Int J Des 6(1):27–39
Pahl G, Beitz W, Feldhusen J, Grote K-H (2007) Engineering design: a systematic approach, 3rd edn. Springer, London
Plattner H, Meinel C, Weinberg U (2009) Design thinking: innovation Lernen – Ideenwelten Öffnen. mi-Wirtschaftsbuch, Munich
Saco RM, Goncalves AP (2008) Service design: an appraisal. Des Manag Rev 19(1):10–19
Sanders EB-N, Stappers PJ (2008) Co-creation and the new landscapes of design. CoDes Int J CoCreat Des Arts 4(1):5–18
Sangiorgi D (2011) Transformative services and transformation design. Int J Des 5(2):29–40

Schön DA (1983) The reflective practitioner: how professionals think in action. Temple Smith, London
Sen A (1999) Development as freedom. Anchor Books, New York
Sleeswijk-Visser F (2009) Bringing the everyday life of people into design. Industrial design. Delft University of Technology, Delft
Tromp N, Hekkert P (2010) A clash of concerns: applying design thinking to social dilemmas. In: DTRS8, Sydney, DAB documents
Tromp N, Hekkert P (2014) Social Implication Design (SID) – a design method to exploit the unique value of the artefact to counteract social problems. In: Proceedings DRS14, Umea
Tromp N, Hekkert P, Verbeek P-P (2011) Design for socially responsible behaviour: a classification of influence based on intended user experience. Des Issues 27(3):3–19
Verbeek P-P (2011) Moralizing technology: understanding and designing the morality of things. University of Chicago Press, Chicago
Vermaas P, Kroes P, van de Poel I, Franssen M, Houkes W (2011) A philosophy of technology: from technical artefacts to sociotechnical systems, vol 6, Synthesis lectures on engineers, technology and society. Morgan & Claypool, San Rafael, CA
Whitney P, Kelkar A (2004) Designing for the base of the pyramid. Des Manag Rev 3:41–47
Winner L (1980) Do artifacts have politics? Daedalus 109:121–136

# Emotions in Design for Values

Pieter M. A. Desmet and Sabine Roeser

## Contents



**Abstract**

The contributions to this handbook show that technology is not value neutral, as is often thought. In this chapter, we argue that the inherent value-ladenness of technology evokes positive and negative emotions of the people who encounter or use it, by touching upon their personal and moral values. These emotions enable people to make concrete practical and moral judgments and to act accordingly. In this chapter, it is therefore proposed that emotions of users and designers alike should not be marginalized as being irrational and irrelevant, but instead be embraced as valuable gateways to values. Emotions reveal those values that matter to our well-being given a particular design or technology, and they are an important source of moral knowledge by being crucial to our capacity of moral reflection. This chapter discusses six sources of emotions in

P.M.A. Desmet (✉) • S. Roeser
TU Delft, Delft, The Netherlands
e-mail: p.m.a.desmet@tudelft.nl; s.roeser@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_6

human-technology interaction and proposes how an understanding of user emotions can support design processes. In addition, the chapter discusses how emotions can resolve the lack of moral considerations in traditional approaches that assess the desirability of technology. It is argued that emotions do this by opening the gateway to moral considerations, such as responsibility, autonomy, risk, justice, and equity. This means that moral emotions can – and should – play an important role in the development of technology and can be considered to be indicators of success and failure in value-driven design processes.

**Keywords**
Emotion • Design • Values • Risk • Well-being

## Introduction

Emotions are usually seen as a distortion of good, rational decision making. In the same vein, emotions might also be seen as a distorting factor in the design of technologies. This chapter challenges this view. Based on recent emotion research, we argue that emotions should play a role in technology design, because emotions reveal important personal and moral values.

The traditional account of technology is that technological design is value neutral and based on rational decision making. According to this account technology is not related to values and emotions because these are seen as a-rational or irrational. There are two challenges to this view. The first challenge is that technology is not value neutral. Technology has impact on our well-being and experiences; it is inherently value-laden. Technology is pervasive: Our daily lives are filled with interactions that are influenced, facilitated, or stimulated by technology. Public services in all domains, such as transport, healthcare, entertainment, and education, heavily rely on technology. Moreover, technology is integrated in all kinds of commonplace consumer products and services, such as telephones, laptops, cars, and dishwashers. People experience positive and negative emotions in response to perceiving, using, and owning consumer products and thus in response to the technology that is integrated in these products. These emotional responses are an expression of personal and moral values and disvalues and can be intended and deliberately designed-for, but they can also be unintended, or even unwanted, and unforeseen by the designer. It is important to already in the design stage explicitly reflect on the values that are affected by technology and to incorporate desirable values in technology and diminish disvalues. We should design for values, an idea which is extensively discussed by the various contributions to this volume. The second challenge to the traditional view of technology is that emotions are not irrational. Recent emotion research has shown that emotions are necessary for our *practical* rationality.

In this chapter, we will show the implications of the combination of these two insights for design theory. We will argue that technologists should consider emotions in the design for values, their own emotions but also those of users and other

stakeholders, as these emotions point out important personal and moral values. We will discuss some ideas on how technology evokes emotions and propose some possibilities for taking these emotions into consideration in the development of new technologies.

## New Approaches to Emotion

Emotions are generally seen as opposed to reason and rationality. Whenever something goes wrong, we blame it on the emotions. When we want things to go right, we invoke rationality. This view is so deeply ingrained in our culture and intellectual heritage that we hardly ever call it into question. In daily language, we (ab)use emotions to explain irresponsible or harmful behavior: "I should not have hit him, but I was blinded with anger," and "I should not have called you in the middle of the night, but I was overwhelmed by the fear of losing you." It is a view that is reflected in empirical decision research, where the dominant theoretical framework is Dual Process Theory. According to Dual Process Theory, we apprehend reality through two different systems: system 1 being emotional and spontaneous and system 2 being rational and reflective. System 1 has the advantage of navigating us smoothly through a complex world but comes at the cost of being highly unreliable. System 2 is normatively superior but requires a lot of time and conscious effort (cf. Kahneman 2011). A similar opposition is prominent in metaethics, the study of the foundations of ethics. The usual taxonomy of metaethical theories consists of sentimentalist versus rationalist approaches to ethics. Sentimentalist approaches see values as expressions of our subjective emotions (Hume 1975 [1748–1752]). Rationalists ban subjective emotions from credible ethical reflection and state that objective values are constituted or understood through rationality (Kant 1956 [1781/1787]). Hence, the dominant approaches to decision theory and value theory endorse the common dichotomy between reason and emotion.

However, this dichotomy has been challenged by recent emotion research. Psychologists and philosophers who study emotions argue that emotions are not opposed to but a specific form of rationality. Emotions are needed in order to be practically rational. The neuropsychologist Antonio Damasio (1994) has studied people with specific brain defects, in the amygdala and in the prefrontal cortex, who don't feel emotions anymore and who at the same time have lost their capacity to make concrete moral and practical judgments. These patients still score equally high on IQ tests as before their illness or accident that caused the damage. They also still know in general that one ought not lie, steal, etc. However, their personality has completely changed. Before their impairment, they were normal, pleasant people, but after their brain damage, they turned into rude people, who in concrete situations are completely clueless on what to do. Hence, emotions turn out to be necessary to make concrete practical and moral judgments and to act accordingly. These ideas are supported by theories from other psychologists and philosophers who emphasize that emotions are not contrary to knowledge and cognition but that

they are themselves a form of cognition and intentionality, so-called cognitive theories of emotions.

A well-accepted cognitive theory of emotions is appraisal theory, which purports that all emotions are elicited by an appraisal (Roseman 1991), an evaluative process that serves to "diagnose" whether a situation has adaptational relevance to the individual and, if so, to identify the nature of that relevance and produce an emotion and an appropriate behavioral response (Lazarus 1991). Someone who is confronted with a fire alarm will most likely experience fear with a corresponding tendency to flee because the fire alarm signals a potentially harmful situation with particular behavioral consequences. This example illustrates that appraisals are inherently relational (e.g., Scherer 1984). Rather than exclusively reflecting either the properties of the stimulus (e.g., a fire), the situation (e.g., the office), or the person (e.g., asthmatic condition), appraisal represents an evaluation of the properties of the stimulus and the situation as it relates to the properties of the individual (Smith and Lazarus 1990). In short, appraisal is an evaluation of the significance of a stimulus for one's personal well-being.

Cognitive theories of emotion especially emphasize the importance of emotions when it comes to our appraisal of personal values. They "pull us toward" ideas, objects, and people that we appraise as favorable and "push us away" from those we appraise as threatening or harmful (Frijda 1986). One's personal values (or "concerns" in the terminology of appraisal theorists, see Frijda 1986) serve as the point of reference in the appraisal process.

These insights from appraisal theorists can shed important light on design for values as follows. An appraisal of designed technology has three basic possible outcomes: the technology is (potentially) beneficial, harmful, or not relevant in relation to our personal values (and thus for personal well-being). These three general outcomes result in a pleasant emotion, an unpleasant emotion, or the absence of an emotion, respectively. Note that in the case of emotional responses to technology, the emotion is not necessarily evoked by the technology itself but can also be elicited by the (imagined, expected, experienced) consequence of the technology or an associated object or person, like the manufacturer or the typical user. Moreover, because appraisal mediates between technology and emotions, the emotion is evoked by the relational meaning of the technology instead of by the technology itself, and different individuals who appraise the same technology in different ways will experience different emotions.

Emotions can reveal personal values and moral values, and the two do not necessarily coincide. One's personal values can be, but are not necessarily, moral values. Indeed, besides moral values, the values that serve as the point of reference in one's emotions can range between values that are morally fully acceptable and values that are morally fully unacceptable. To take an extreme example, a hunter who enjoys hunting for endangered species like elephants may experience positive emotions in his actions because these match his personal values ("freedom to hunt"), even though other people may feel that these are morally intolerable.

Even though the hunter's pleasure reflects a personal value that is not a moral value, he may experience additional emotions in relation to his activity that do serve

moral functions. For example, if his friends or family respond with contempt instead of pride to his hunting trophies, he could feel embarrassed, and this emotion can stimulate him to rethink the moral qualities of his hobby. This example illustrates that emotions can not only result from cognitions, as is emphasized by appraisal scholars, but they can themselves be a source of cognition. According to some forms of cognitive theories of emotion, emotions are affective and cognitive at the same time (Zagzebski 2003; Roberts 2003). Emotions let us see the world in a specific light and let us focus on morally salient features (Little 1995). Emotions draw our attention to what matters, in our own lives but also in those of other people (Little 1995; Blum 1994). By drawing our attention to our personal values, emotions can stimulate us to reflect on the moral implications of these values. For example, emotions like guilt can increase our awareness of how our actions conflict with moral values. Likewise, such emotions can make us aware that a personal value may be morally unacceptable (Camacho et al. 2003). Social emotions such as compassion help us to extend our "circle of concern" from near and dear ones to people far away (Nussbaum 2001). Feelings of sympathy, responsibility, and care help us to understand that we should help others and what their needs might be. Emotions are an important source of moral knowledge, understanding, and awareness (Roeser 2011).

## The Role of Emotions in the Experience and Evaluation of Technology

If we combine the insight that technology is value laden with the insight that emotions are a prime source of knowledge and understanding of values, it follows that emotions can and should play an important role in understanding values involved in technology.

Technology can affect our well-being, for better or worse. Traditional approaches to assess the desirability of a technology are based on risk-benefit analysis. According to such an approach, the benefits of a technology are measured in, for example, economic terms and balanced against possible negative side effects or risks. Risk is defined as the probability of an unwanted effect. This approach to risk has been severely criticized by social scientists and philosophers, who have pointed out that this approach is too narrow (Slovic 2000; Krimsky and Golding 1992). It is difficult or even impossible to express all moral considerations about technologies in terms of risks or costs and benefits and to compare them on one scale (Espinoza 2009). Risk-cost-benefit analysis leaves out important ethical considerations such as responsibility, autonomy, justice, fairness, and equity (Asveld and Roeser 2009). Moral emotions related to risk such as indignation, compassion, and feelings of responsibility can point out such moral considerations that cannot be captured in a traditional risk-cost-benefit analysis (Roeser 2006, 2010).

Conventional approaches to risk assessment leave out important values, but they also ignore emotions, as they are seen as a threat to rational decision

making about technologies. Even scholars who emphasize the importance of a broader perspective to risk, including moral values, struggle with the role emotions might play in assessing risky technologies. Based on Dual Process Theory, Slovic et al. (2004) think that risk emotions should be corrected by quantitative methods. Loewenstein et al. (2001) argue that the emotions of the public have to be respected simply because we live in a democratic society, even though they might be irrational. Sunstein (2005) argues that we should even avoid emotions in risk judgments and use cost-benefit analysis instead.

However, as argued before, emotions should not be seen as contrary to rationality, as Dual Process Theory has it, but rather, they should be seen as a form of practical rationality. This idea can shed completely new light on risk emotions. They are not an obstacle to decision making about risky technologies; rather, they are a source of awareness of moral values that are involved in risky technologies. Risk and value-sensitive design can be seen as two sides of the same coin. With value-sensitive design, we try to diminish the potentially negative effects of risky technologies. Emotions such as sympathy, compassion, indignation, and feelings of responsibility allow us to be sensitive to ethical aspects of technologies such as justice, fairness, equity, and autonomy. This awareness is an important first step in critically reflecting about the kinds of values that we want to be included in the design of a technology.

As Papanek (1985) already stressed in his famous book *Design for the Real World*, design has the ability to create well-being, it can embody the principles of good citizenship, and it can challenge, engage, and nourish culture and identity. Over the last few years, a growing group of designers and engineers in both industry and academia has been inspired by the possibility to increase the subjective well-being of individuals and communities. Seligman and Csikszentmihalyi (2000), p. 5, purport that the social and behavioral sciences can play an enormously important role in nurturing human welfare: "They can articulate a vision of the good life that is empirically sound while being understandable and attractive. They can show what actions lead to well-being, to positive individuals and to thriving communities." We propose that in line with this thought, the design discipline can play an equally important role by materializing the vision of the good life, enabling and stimulating actions that lead to well-being and thriving communities (cf. Desmet et al. 2013). But then emotions should play an important role in design, as they help us to draw our attention to what matters to our own well-being and that of others.

Because emotions can facilitate and stimulate but also discourage or obstruct technology usage, they can, do, and should play an important role in the process of developing technology. In product design, measuring emotions elicited by existing products has been proven an adequate means for uncovering relevant (and often unanticipated) values that drive the emotional responses of users toward existing products. These insights can be used to formulate value profiles that direct new technology development. An example is a wheelchair design for children (see Desmet and Dijkhuis 2003). The emotions experienced by children in response to existing wheelchairs were measured, and negative emotions served as cues that the design threatened user values. One of the findings was that children experience

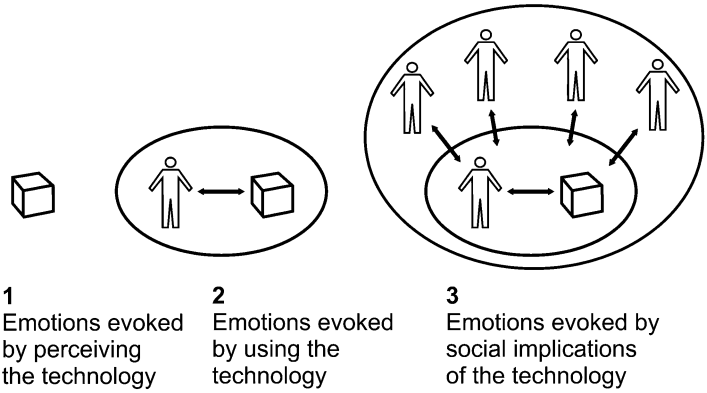



**Fig. 1** Three main sources of technology emotions (Adapted from Desmet 2012)

contempt in response to wheelchairs with big handles. Interviews revealed that the cause of this emotion was that the big handles conflict with the personal value of "being independent" (i.e., having big push handles expresses dependency). This was a relevant insight for a redesign in which the children can freely slide the handle behind the back side when using the wheelchair individually. By not being recognizable, the handle no longer expresses dependency. This example illustrates that emotions can operate as portals to relevant personal values. User emotions should therefore play an important role in design processes, and because these emotions are valuable sources of moral knowledge, they should be taken into consideration in the evaluation of ethical aspects of technology. Note that technology often evokes mixed emotions because emotions are evoked by different levels of interaction. Figure 1 visualizes three main sources of emotions experienced in relation to designed technology (Desmet 2008, 2012). The first represents emotions experienced when perceiving (seeing, touching, tasting, thinking or hearing about, etc.) technology; these emotions are "about" the design of the technology as such. The second represents emotions experienced when using technology to fulfill its purpose; these emotions are "about" the activity of using the technology. The third represents emotions experienced in relation to the social implications of using and owning the technology; these are emotions "about" one's relationship with other people. Below, we discuss these three sources with the intention to illustrate that technology tends to have a multifaceted rather than a single emotional impact. This multifaceted nature is particularly interesting because each of the sources can evoke emotions that are an expression of moral values and/or a source of moral knowledge.

Each of these three sources can be detailed in at least two subordinate sources. Table 1 gives an overview of the various sub-sources of emotions, with a simple black pen as an example. The sources differ in terms of the trigger cause (or focus) of the emotion: the emotion can be evoked by (1) the design of the technology as such, (2) the symbolic or associated meanings, (3) the behavior of the technology when in use, (4) the activities that are influenced or enabled by using the

**Table 1** Emotions in response to design (or designed technology) (Adapted from Desmet 2012)

| Perceiving the design (or designed technology) | Using the design (or designed technology) | Social implications of (using or owning) the design (or designed technology) |
|---|---|---|
| (A) Object-focus | (C) Usage-focus | (E) Relationship-focus |
| Emotions evoked by the material qualities of the design | Emotions evoked by the interactive qualities when using the design | Emotions evoked by the influence of the design on one's relationships with other people |
| What do you see when looking at the design? | How does the design respond to you when using it? | What effects does the design have on your social relationships? |
| "I enjoy looking at this unique pen that is made of sustainable bamboo" | "I enjoy the ease of using this pen because the weight distribution is perfectly balanced" | "I feel reluctant when people ask me if they can borrow my pen because it is fragile and I would not like it to be damaged" |
| (B) Association-focus | (D) Activity-focus | (F) Identity-focus |
| Emotions evoked by something (or someone) that is represented by the design | Emotions evoked by the consequences of using the design | Emotions evoked by the influence of using or owning the design on one's social identity |
| What do you know about the design? | What does the design enable you to do? | What does owning or using the design say about you? |
| "I cherish this pen because it represents my passion for the combination of beauty and sustainability" | "Drawing is an activity that makes me energetic" | "I am proud of being person who takes good care of his belongings" |

technology, (5) the social implications of using the technology, and (6) the impact of using or owning the design on one's personal identity.

## Emotions Evoked by Perceiving Design

Emotions can be evoked by the perceivable manifestations of technology. An individual can, for example, love an advanced computer for its beautiful design. Or one can be curious about a novel design or fascinated by a complicated design or feel sympathetic toward broken-down or obsolete technology. Appearance is used in the broad sense of the word, involving not only the visual appearance but also the taste, tactile quality, sounds, and fragrances.

Sometimes, the emotion is not directly evoked by the technology's appearance but by some associated object, person, event, or belief. One can, for example, admire the manufacturer of an innovative technology (in this case the object of the emotion is the manufacturer) or love a design because it reminds one of a loved person (in this case the object of the emotion is the loved person). Designed technology often represents or symbolizes intangible personal (moral) values or beliefs. Some products are deliberately designed to represent such values or beliefs.

Examples are spiritual and religious objects, tokens, mementos, souvenirs, keepsakes, talismans, and mascots. In other cases, technology is not intentionally created to represent values or beliefs and obtains its symbolic value in user-technology interaction or in the cultural discourse. Note that a special type of emotions evoked by technology are those that are related to anticipated usage or anticipated consequences of usage. When being introduced to new technology, people anticipate on how it will be to use or benefit from this technology. Or one can experience hope in response to a mobile phone because one anticipates that it will support one's social life or fear in response to a new technology to produce energy because one anticipates that it will harm their moral norm of being energy efficient.

## Emotions Evoked by Using Technology

We use technology with the purpose to fulfill needs or achieve goals. This can be to drill a hole in a wall, to listen to music, to cook a meal, etc. The activity of using technology can evoke various kinds of emotions. These can be emotions evoked by the interaction with the technology as such (usage-focus) or by the activity that is enabled or facilitated by using the technology (activity-focus). In the first case, the emotion is evoked by how the technology responds to us when using it. For example, the technology can be easy to use or complicated and challenging. It can behave unexpectedly or predictably. This "quality of interaction" can evoke all kinds of emotions. One can become energetic by technology that requires physical effort to use and experience joy when technology is unexpectedly easy to use or pride when successfully operating complicated technology.

In the second case, the emotion is evoked by the activity that we engage in when using technology. Technology is used to enable or facilitate all kinds of activities; it provides us with instruments that are used to "get something done" in some situation: activities that can be useful (e.g., organizing my documents) or pleasurable (e.g., ice-skating with friends) or morally commendable (e.g., helping out a neighbor). Individuals will respond emotionally to these activities because they have personal values related to the activities. The emotion is not directed to the technology as such, but the technology does play a role because it enables the individual to engage in the activity that evokes the emotion. Examples are I am excited by making a hiking trip in the snow (which is facilitated by my GPS system to keep me safe); I enjoy talking to my friends (which is facilitated by my mobile phone); and I am satisfied with the stack of clean laundry (which is facilitated by my washing machine).

In many cases, users do not have a direct emotional intention when using designed technology. In those cases, emotions are "side effects," like unexpected sensorial pleasures of using the technology. In other cases users do have a deliberate intention to affect their emotions when using technology. Examples are computer games and relaxing chairs. We use computer games because they amuse us, and we use relaxing chairs because they relax us and ride a motorcycle because it excites us.

## Emotions Evoked by the Social Implications of Technology

Technology is always used in some social context. We use technology in our interactions with other people (e.g., communication devices and gifts), and the technology that we use and own affects our social identity. In the first case, the emotion is evoked by social interactions that are influenced or facilitated by the technology. One can enjoy talking to a friend (facilitated by a phone), be proud of being able to help someone (facilitated by a city map on one's smartphone), or enjoy drinking a glass of wine with a group of friends (facilitated by an online event planner).

In the second case, the emotion is evoked by our identity, as affected by using or owning technology. As was mentioned by Belk (1988), products are extensions of their owners, and they affect an individual's self-perception and how they are perceived by others. An expensive stroller might support someone in their self-perception of being a good parent, crayons enable someone to be a creative person, and an SUV car makes someone look cunning or irresponsible, depending on one's personal values. People are emotional about who they are and how others perceive them and thus also about the effects of the technology they use and own on their identity. Examples are I feel insecure because I have to use hearing aids (which conflicts with my personal value of being independent) and I feel confident when driving my new electric car (which matches with my moral value of not wasting fossil energy).

## Design for Values

The insights on how emotions play a role for the users of technology can help the developers of a technology in their design. In line with the three sources of emotions elicited by technology, we propose that there are three levels of emotional appeal: an object appeal (the degree to which the technology is appealing), an activity appeal (the degree to which my activity is appealing), and a self-appeal (the degree to which I am, or my life is, appealing). In this context, the word "appealing" is used for technology that is appraised as beneficial to our personal values. To design for each level of appeal requires different considerations because different values will be involved. Personal values and emotions are interrelated: Emotions experienced by users reveal the personal values of these users, and when designers have the intention to deliberately design for particular emotions, they should have an overview of the users' personal values that can be affected by the design. This means that emotion-driven design is actually value-driven design.

The first step is to identify the user group and the situation in which the technology will be used. This can be formulated in the form of a design theme, expressing a user group engaged in an activity in some situation. Examples are police officers using a communication device when at work in the streets or caretakers using a bottle when feeding a toddler at home. The second step is to

formulate a value profile that represents this design theme. Because many personal values can be at stake in a given situation, the challenge is not to aim for completeness but for a concise value profile that is both relevant and inspiring. In line with the three levels of appeal, the profile includes values related to the technology itself, to the activity facilitated by the technology, and to the social impact of using or owning the technology. Key questions in formulating values related to the technology itself are: what are the users' expectations and standards about this technology, and what kind of properties do they enjoy? Examples are a phone should be strong enough not to break when I drop it, and I want my table to be made of honest materials. Key questions in formulating activity values are: what do the users want to accomplish in the usage situation, and what do they expect from themselves? Examples are I should be patient with my clients at work, and I want my son to enjoy himself at school. Key questions in selecting values related to the social impact of the technology are: what do the users expect from themselves in life, and what are the general life goals and aspirations they pursue? Examples are I should be fit and I want to be autonomous.

Once the value profile is defined, a design profile can be formulated that represents the designer's vision on how to align with the value profile, specifying three qualities: the product's significance, intentions, and character. Significance represents the key consequences that we want to design for; e.g., I have many friends, I am relaxed, my baby is happy, or I am inspired. The intentions represent the purpose it will be designed to have, such as the technology enables me to talk freely and enables me to meet people, to be spontaneous, or to be at work on time. The character represents the technology's appearance, such as the design is rough, inviting, delicate, natural, or colorful. The design profile is used to formulate a product statement. Some examples are a delicate product that enables me to have a relaxed life by seducing me to talk freely and a tough product that enables me to have many friends by forcing me to open up to others. The value profile and product profile can be used as a reference in all stages of the design process in order to safeguard the emotional fittingness of the final design.

## The Role of Emotions in Design for Values

So far we have discussed how designers can incorporate the personal values and emotions of the users. However, the emotions and personal values of the users or clients might be morally contentious, or these users might not be aware of the potentially morally problematic implications of a technology. As was mentioned earlier, not all personal values are necessarily moral values, and some can even be morally unacceptable. This means that it is necessary to include critical reflection on these emotions and values in order to make sure that we do not design for any values, but for moral values, or at least morally acceptable values. Here, however, emotions can also play an important role, as they endow us with the capacity to critical moral reflection (Roeser 2010). Emotions such as compassion and feelings

of responsibility can entice us to counteract selfish emotions. For example, our care for the environment can entice us to design, buy, and use a more sustainable but slower car (Roeser 2012). Several methods have been developed to enable reflection about technology, for example, scenarios that describe situations in which the use of a technology gives rise to moral considerations (cf. Boenink et al. 2010). These methods involve narratives that directly engage the imaginative capacities of people. These methods can be further developed to explicitly encourage emotional engagement and emotional reflection. This enables critical reflection about one's own lifestyle and considerations of justice toward others. By providing people with concrete narratives, distant others that can otherwise easily be neglected come uncomfortably close by and force oneself to critically assess one's own behavior.

However, emotions can play yet another role in technology design. The developers of the technology themselves presumably have emotional responses to their designs. They should take these emotions seriously as they can draw their attention to important values that can be potentially incorporated in the design. These can be positive values, that should be maximized, and negative values, that should be minimized. The designers can oversee and influence the properties of a technology more directly than anybody else, which gives them a special responsibility (Van der Burg and Van Gorp 2005). Experts might be concerned about unpredictable consequences of a technology. However, even if the consequences of a technology are fairly well known, there can be remaining emotional-ethical concerns that should be taken seriously, such as potential misuse of a technology or potentially or even explicitly immoral requirements set by the client or user. Designers should use their imaginative capacities, for example, by empathizing with possible victims of a technology, in order to come to a more active appreciation of their moral responsibility in designing risky technologies. Designers can take on stronger responsibilities if they cherish their imaginative, emotional capacities. This will make them feel more involved, responsible, and prone to take action. Designers should take on this responsibility to which their emotions can draw their attention (Roeser 2012). Drawing on the reflective, critical capacity of emotions can make an important contribution to design for moral values.

## Future Research

In the section "Design for Values", we have discussed design for emotional experience; in sections "The Role of Emotions in the Experience and Evaluation of Technology" and "The Role of Emotions in Design for Values", we have discussed emotional evaluation of technology. Now these perspectives can be combined in future research, through the idea of reflective technologies, in other words, technology as a means for meaningful activities. The idea is that technologies themselves can give rise to, entice, and encourage critical reflection on what are desirable activities enabled through technology. Here emotions and values play an important role again. On the conventional view of emotions as opposed to rationality, design that appeals to

emotion entails evoking unreflected gut reactions. However, based on the novel theories of emotions sketched in the section "New Approaches to Emotion", appealing to emotions through design can endow us with the capacity to take a critical stance toward a technology and the kind of behavior it invokes.

Let us first take a look at how technological design might entice critical reflection about our personal values. A lot of work has been done recently on the way technological, institutional, and other designs can "nudge" people to do certain things that are ethically desirable (Thaler and Sunstein 2008). However, nudging might lead to manipulation. One might argue that as long as it is for a greater good, manipulation is justified. However, this is a very consequentialist way of reasoning that is ethically dubious, as it might not respect the autonomy and reflective capacities of people. Thaler and Sunstein (2008) argue that manipulation cannot be avoided. Any presentation of options steers our choices and behavior. Based on this, they argue that choice options ("nudges") should be provided that let us do things that we would endorse. However, technology design does not only need to work as simple nudge, but it can also be a vehicle for reflection.

Indeed, the design discipline has a rich tradition in using design as an instrument to stimulate discussion and reflection. In the 1960s and 1970s, for example, designers and architects in the Italian "Radical Design" movement used design to embody their critical views on prevailing material culture and technology values. More recently, Dunne and Raby (2001) proposed the concept of "design noir" as a reaction to the (in their view) impoverished experiential value of mainstream consumer technology. Design noir was offered as a new genre of design to explore how technology can be designed that expands our experience of everyday life. Using design as a means for reflection was coined "critical design" in 1999 by Dunne in his book *Hertzian Tales*. Critical design aims to provoke users in reflecting on their values and practices by challenging preconceptions and expectations. In that way, critical design can stimulate new ways of thinking about technology, its usage and meaning.

Emotions can play a role in this critical design, by, for example, making users feel uncomfortable while using a certain technological product, presenting them with surprising, disgusting, and frightening experiences that force them to reflect on their behavior, value patterns, and responsibilities. Demir (2010) described "Poor Little Fish" designed by Yan Lu (Fig. 2; see www.yanly.com) as an example of such design. The product combines a fish bowl and a water tap and challenges people to reflect on how their behavior touches upon their personal values of sustainability.

While using the tap, the level of the water in the bowl gradually falls (but does not actually drain out); it will return to the original level once the water stops running. The combination of a tap and a fish tank draws a parallel between water consumption and damage to natural life. Emotions such as sympathy for the fish, the fear of killing it, or shame to disturb its home can stimulate a direct tendency to reduce water consumption and a more indirect behavioral effect mediated by reflection. Although critical design generally provokes an unfavorable view on the existing role of technology in our daily lives, we believe that design can also

**Fig. 2** "Poor Little Fish" water tap, by Yan Lu

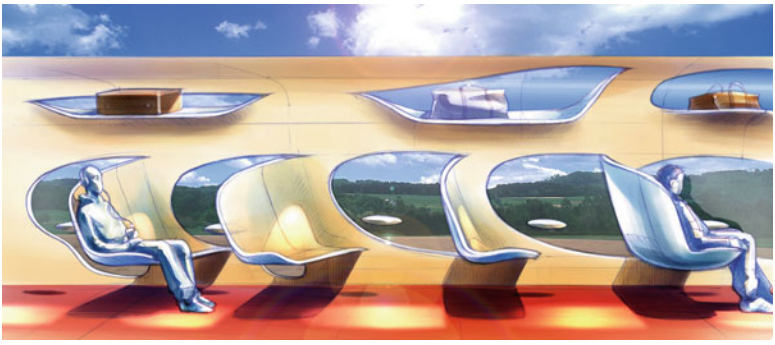


**Fig. 3** Connecting Europe, concept high-speed train by Doeke de Walle

be used to stimulate discussion and reflection on opportunities and possibilities of newly conceived technology. This "constructive design," which embodies manifestations of future technologies, can be provocative too, expressing possibilities previously thought not realizable and stimulating discussions on future rather than current material culture and technology values. Here again is an important possible role for emotions. Constructive design can trigger our imagination and compassion and endow us with inspiration and motivation to try something new that might make a difference. Desmet and Schifferstein (2011) presented a collection of 35 experience-driven design projects that illustrate the inspirational quality of constructive design. An example is a concept for a trans-European high-speed train, designed by Doeke de Walle for Pininfarina; see Fig. 3.

The train was designed to express new possibilities of novel layered construction methods and materials. This enabled a design that offers unobstructed views of the

external surroundings, connecting the outside to the inside world, which stimulates the value of "freedom of movement" and evokes emotions like anticipation and delight.

We can think of the role that social media such as Facebook have played in recent political movements against oppressive regimes, as a platform for sharing both negative and positive emotions, such as frustration, anger, hope, and relief. Social media enable people to build relations and feelings of connectedness with a large number of people whom they cannot reach easily via other means. This can endow them not only with the practical tools to reach out to large groups but also to build a feeling of community, trust, and shared interests. Future research should investigate the many possible ways in which emotions can play a role in reflective design. The frameworks sketched earlier on in this chapter, i.e., cognitive theories of emotions and emotion-driven design, can provide for a basis from where to explore these possibilities.

## Conclusion

Emotions should play an important role in design for values because the emotions that people experience in response to design and technology are an expression of their personal and moral values. In other words, emotions are the gateways to value: We are only emotional about things that touch upon our personal and moral values. This implies that negative emotions are just as relevant as positive emotions, because both indicate underlying values. We can distinguish different layers of emotions in response to technology. When aiming to understand an emotional response, we should be aware that this emotion can be evoked by the technology itself, but also by activities that are enabled and supported by the technology, or by the impact of technology on one's social identity. The ability of designing technology that evokes positive (and prevents negative) emotions can be increased by formulating value profiles that represent a particular user group and a particular usage situation; it is a combination of general values and contextualized values that drive emotional responses to technology in everyday life. Technology and design have an enormous potential for promoting well-being. Emotions can play an important role in pointing out the values that matter for our well-being and that of others. Emotions are elicited by all aspects of design and technology that are perceived as good or bad, desirable or distasteful, effective or useless, and meaningful or pointless – they are both an expression of personal and moral value and an entry point to these values. That is why, rather than being a distorting factor in the design of technologies or the "cherry on the cake," a finishing touch that is added to a design that has already been optimized on all other aspects, emotions should be considered to be a valuable source of information and indicator of success and failure in any value-driven design process. Our emotions reveal what we value, to ourselves, to the people we encounter, and ideally also to those who design the technologies that we live with.

## Cross-References

- ▶ Conflicting Values in Design for Values
- ▶ Design for the Value of Human Well-Being
- ▶ Design for the Value of Safety
- ▶ Design for the Value of Sustainability
- ▶ Design for the Values of Democracy and Justice
- ▶ Design for Values in Engineering
- ▶ Design Methods in Design for Values
- ▶ Participatory Design and Design for Values

## References

Asveld L, Roeser S (2009) The ethics of technological risk, risk, society and policy series. Earthscan, London

Belk RW (1988) Possessions and the extended self. J Consum Res 15(2):139–168

Blum LA (1994) Moral perception and particularity. Cambridge University Press, Cambridge/New York

Boenink M, Swierstra T, Stemerding D (2010) Anticipating the interaction between technology and morality: a scenario study of experimenting with humans in bionanotechnology. Stud Ethics Law Technol 4(2):1–38

Camacho CJ, Higgins ET, Luger L (2003) Moral value transfer from regulatory fit: what feels right is right and what feels wrong is wrong. J Pers Soc Psychol 84(3):498–510

Damasio AR (1994) Descartes' error: emotion, reason and the human brain. G.P. Putnam, New York

Demir E (2010) Understanding and designing for emotions. Unpublished PhD dissertation, Delft University of Technology, Delft

Desmet PMA (2008) Product emotion. In: Hekkert P, Schifferstein HNJ (eds) Product experience. Elsevier, Amsterdam, pp 379–397

Desmet PMA (2012) Faces of product pleasure: 25 positive emotions in human-product interactions. Int J Des 6(2):1–29

Desmet PMA, Dijkhuis EA (2003) Wheelchairs can be fun: a case of emotion-driven design. In: Proceedings of the international conference on designing pleasurable products and interfaces, ACM, Pittsburgh/New York, 23–26 Jun 2003

Desmet PMA, Schifferstein HNJ (2011) From floating wheelchairs to mobile car parks: a collection of 35 experience-driven design projects. Eleven Publishers, Den Haag

Desmet PMA, Pohlmeyer A, Forlizzi J (eds) (2013) Special issue on design for subjective well-being. Int J Des 7(3):1–3

Dunne A (1999) Hertzian tales; electronic products, aesthetic experience, and critical design. MIT Press, Cambridge

Dunne A, Raby F (2001) Design noir: the secret life of electronic objects. Berkhauser, Berlin

Espinoza N (2009) Incommensurability: the failure to compare risks. In: Asveld L, Roeser S (eds) The ethics of technological risk. Earthscan, London

Frijda NH (1986) The emotions, studies in emotion and social interaction. Cambridge University Press, Cambridge

Hume D (1975 [1748–1752]) Enquiries concerning human understanding and concerning the principles of morals. Edited by LA Selby-Bigge. Clarendon Press, Oxford

Kahneman D (2011) Thinking fast and slow. Farrar, Straus and Giroux, New York

Kant I (1956 [1781/1787]) Critique of practical reason (trans: White Beck L). Bobbs-Merrill, New York

Krimsky S, Golding D (1992) Social theories of risk. Praeger Publishers, Westport
Lazarus RS (1991) Progress on a cognitive motivational relational theory of emotion. Am Psychol 46(8):819–834
Little MO (1995) Seeing and caring: the role of affect in feminist moral epistemology. Hypatia J Fem Philos 10(3):117–137
Loewenstein GF, Weber EU, Hsee CK, Welch N (2001) Risk as feelings. Psychol Bull 127:267–286
Nussbaum MC (2001) Upheavals of thought: the intelligence of emotions. Cambridge University Press, Cambridge
Papanek VJ (1985) Design for the real world: human ecology and social change, 2nd, completely rev. edition. Thames and Hudson, London
Roberts RC (2003) Emotions: an essay in aid of moral psychology. Cambridge University Press, Cambridge/New York
Roeser S (2006) The role of emotions in judging the moral acceptability of risks. Safety Science 44:689–700
Roeser S (ed) (2010) Emotions and risky technologies. Springer, Dordrecht
Roeser S (2011) Moral emotions and intuitions. Palgrave Macmillan, Basingstoke
Roeser S (2012) Emotional engineers: toward morally responsible engineering. Sci Eng Ethics 18(1):103–115
Roseman IJ (1991) Appraisal determinants of discrete emotions. Cogn Emot 5(3):161–200
Scherer KR (1984) On the nature and function of emotion: a component process approach. In: Scherer KR, Ekman P (eds) Approaches to emotion. Erlbaum, Hillsdale
Seligman EP, Csikszentmihalyi M (2000) Positive psychology; an introduction. Am Psychol 55 (1):5–14
Slovic P (2000) The perception of risk, risk, society, and policy series. Earthscan, London/Sterling
Slovic P, Finucane ML, Peters E, Macgregor DG (2004) Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. Risk Anal Int J 24(2):311–322
Smith CA, Lazarus RS (1990) Emotion and adaptation. In: Pervin LA (ed) Handbook of personality: theory and research. Guilford, New York
Sunstein CR (2005) Laws of fear: beyond the precautionary principle. John Robert Seeley lectures. Cambridge University Press, Cambridge
Thaler RH, Sunstein CR (2008) Nudge: improving decisions about health, wealth and happiness. Penguin, London
Van der Burg S, Van Gorp A (2005) Understanding moral responsibility in the design of trailers. Sci Eng Ethics 11:235–256
Zagzebski L (2003) Emotion and moral judgment. Philos Phenomenol Res 66:104–124

# Human Capabilities in Design for Values

## A Capability Approach of "Design for Values"

Ilse Oosterlaken

## Contents



**Abstract**

Technology and the expansion of human capabilities are intimately related. This chapter discusses an influential philosophical framework that attaches central moral importance to human capabilities, namely, the so-called capability approach, and explains in which ways it has relevance for design. A distinction will be drawn between two different, although related, design applications of the capability approach. Firstly, in the "narrow" usage, the capability approach is seen as presenting a proper conceptualization of individual well-being, namely, in terms of the capabilities that a person has. The aim of design is then to contribute to the expansion of these capabilities, to which I refer as design for capabilities.

I. Oosterlaken (✉)
Department of Values and Technology, Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands
e-mail: e.t.oosterlaken@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_7

I will discuss two challenges for design for capabilities, namely, an epistemological and an aggregation challenge. Secondly, in the "broad" usage, the capability approach is seen as a source of insight and inspiration for taking a broader range of values and concerns into account in design, most importantly agency and justice. From this perspective, so it is argued, strong parallels can be drawn with participatory design and universal design. In reality both the narrow and the broad usage of the capability approach in design should go hand in hand. The chapter ends with some reflections on the challenges ahead in making the philosophical literature on the capability approach accessible to and usable by designers.

**Keywords**
Agency • Justice • Well-being • Capability approach

## Introduction

As mundane as some technological artifacts may seem to be, there is sometimes a rich story to be told about their meaning for or impact on human lives. Take, for example, a lamp. It has a rather straightforward function: to give light. Since lamps are ubiquitous in modern, Western society, we rarely stop to reflect on it. Yet due to factors such as low income or the absence of an electricity infrastructure, having light is not self-evident for everyone. In 2008 I met an industrial design engineer who had worked on several design projects for poor communities in the South, including the design of lamps. The experiences gained during that work, so he told me, made him realize that lamps are ultimately not about light. The importance of a lamp lies in the fact that it enables you to do things that contribute to the overall quality of life, for example, to go to the outdoor toilet at night without being afraid or to make your homework in the evening after having looked after your family's cattle all day. Technology has, so this simple example illustrates, the potential to contribute to the quality of life by expanding what people can do or be – their capabilities.

That technical artifacts have in essence something to do with enabling human action, with expanding human capabilities, is an intuitively plausible idea that has recently been reflected upon by several philosophers of technology (e.g., Lawson 2010; Van den Hoven 2012; Illies and Meijers forthcoming). The focus of this chapter will, however, be on a more general philosophical framework that attaches central *moral importance* to certain human capabilities, namely, the so-called capability approach. In this approach – for which Amartya Sen and Martha Nussbaum have done much of the groundwork – human capabilities are often described as the real opportunities for a person to do and be what he/she has reason to value. In a recent introduction to the capability approach, Robeyns (2011) notes that

> it is generally understood as a conceptual framework for a range of normative exercises, including most prominent the following: (1) the assessment of individual well-being; (2) the evaluation and assessment of social arrangements; and (3) the design of policies and proposals about social change in society. (Robeyns 2011, p. 3)

This chapter will discuss the capability approach as a normative framework that also has relevance for design – be it engineering design, industrial design, or architectural design. Furthermore, it should perhaps be noted at this point that I am rather lenient toward what counts as "design." This could be conceptualizing and shaping a completely new artifact/system, redesigning and improving an existing artifact/system, or merely trying to figure out the best configuration of an artifact/system based on existing components and technologies. The chapter can be seen as providing a specific elaboration of the general idea of "design for values" that is the central topic of this handbook.

The structure of this chapter is as follows. I will start with an outline of the central concepts and philosophical ideas present in the capability approach (section "The Capability Approach").[1] It will then be briefly discussed how technology and human capabilities are related (section "The Complex Relation Between Technology and Human Capabilities"). The first two sections thus provide the background against which the remainder of the chapter explores in more detail the different ways in which the capability approach could be relevant to design. A distinction will be drawn between the two usages of the capability approach.[2] In the "narrow" usage[3] (section "The "Narrow" Application of the Capability Approach: Well-Being"), the capability approach is seen as presenting a proper conceptualization of individual well-being, with the aim of design being able to contribute to this. This, however, raises some discussion points and a number of problems, most importantly an epistemological challenge

[1]This section should give designers a minimal basis for the "conceptual investigation phase" of the tripartite "value sensitive design" or VSD approach developed by Friedman and her colleagues (e.g., Friedman et al. 2001; see also Chap. 2 of this book). According to the VSD approach, these conceptual investigations should be closely intertwined with empirical and technical investigations throughout the design process. In that light, it could be considered an attractive feature of the capability approach that – in addition to the philosophical literature – there also exists a large and interdisciplinary body of literature discussing its "operationalization" and presenting empirical applications. This social science literature, although not further discussed in this chapter, may be relevant for designers in two ways. Firstly, the methodologies used to evaluate well-being and social arrangements in terms of human capabilities may also be useful for the evaluation of design alternatives or final design outcomes. Secondly, the results of such empirical studies may be useful, by providing designers with relevant knowledge about (a) stakeholder views on which human capabilities are important and how they should be understood and (b) factors contributing to or inhibiting the expansion of human capabilities in concrete contexts of usage.

[2]Both are already referred to implicitly in my article in *Design Issues* (Oosterlaken 2009), which talks about design that aims to expand human capabilities and also links the idea of "capability sensitive design" to participatory design and universal/inclusive design. Yet the explicit distinction made in this chapter was not made in that article.

[3]"Narrow" should not be read as implying a value judgment. See Robeyns (2011) for an explanation of the distinction between a narrow and a broad employment of the capability approach. She contrasts the broad usage in two different ways with the narrow usage (a) taking into consideration a broader range of values versus being concerned with well-being alone and (b) focusing on the evaluation of policies and social institutions vs. focusing on the well-being of individuals. I'm using distinction (a), applied to the normative evaluation of design, so comparable to the evaluation of policies and institutions in distinction (b).

(section "The Epistemological Challenge for a Well-Being Application") and an aggregation challenge (section "The Aggregation Challenge for a Well-Being Application"). In the "broad" usage, the capability approach is seen as a source of insight and inspiration for taking a broader range of values and concerns into account in design – most importantly agency (section "A "Broad" Application of the Capability Approach: Agency") and justice (section "A "Broad" Application of the Capability Approach: Justice").[4] I will next discuss some general challenges for a capability approach of design (section "Looking Ahead: Some Further Challenges") and end with some brief conclusions.

## The Capability Approach

One way to view the capability approach is as a position in the debate about the best "informational basis" for judgments about justice, equality, well-being, and development. According to the capability approach, assessment should not primarily take place in terms of income, resources, primary goods, utility (i.e., happiness or the sum of pains and pleasures), or preference satisfaction. The focus should rather be on a range of human capabilities. These capabilities are generally described as what people are *effectively* able to do and be or the positive freedom that people have to enjoy valuable "beings and doings." These beings and doings are called "functionings" by Sen. Examples of functionings are such diverse things as working, resting, being literate, being healthy, being part of a community, being able to travel, and being confident. Functionings "together constitute what makes a life valuable" (Robeyns 2005) and are "constitutive of a person's being" (Alkire 2005a). "The distinction between achieved functionings and capabilities," so Robeyns (2005) explains, "is between the realized and the effectively possible; in other words, between achievements on the one hand, and freedoms or valuable options from which one can choose on the other."

As Alkire explains, one reason to focus on capabilities instead of functionings is that we value free choice and human agency. "Agency," so Alkire (2005b) explains, "refers to a person's ability to pursue and realize goals that he or she values and has reason to value. An agent is 'someone who acts and brings about change.' The opposite of a person with agency is someone who is forced, oppressed, or passive." Nussbaum (2000) conceptualizes the human being as "a dignified free being who shapes his or her own life"; she says, "we see the person as having activity, goals, and projects." The idea is that if people have a range of different capabilities, they may choose to realize those functionings that are in line with their view of the good life. Policies should – according to the capability approach – aim at expanding people's capabilities and not force people into certain functionings.

[4]This means that there will be commonalities with some of the other chapters in this handbook, such as that on "▶ Design for the Value of Human Well-Being," "▶ Design for the Values of Democracy and Justice," and "▶ Design for the Value of Inclusiveness."

"The 'good life' is partly a life of genuine choice," says Sen (1985), "and not one in which the person is forced into a particular life – however rich it might be in other respects." One thing is certain: as a range of capabilities is always needed for a good human life, well-being is multidimensional according to the capability approach. This, so Nussbaum (2000, p. 81) has argued, "limits the trade-offs that it will be reasonable to make."[5] If someone lacks, for example, the capability to be well nourished, we cannot – or at least not fully – compensate this deprivation by expanding his capability to maintain meaningful social relations.[6]

Why should we focus on capabilities rather than utility or resources? A main reason is that the relationship between a certain amount of goods and what a person can do or can be varies, as Sen and others have often illustrated:

> . . . a person may have more income and more nutritional intake than another person, but less freedom to live a well-nourished existence because of a higher basal metabolic rate, greater vulnerability to parasitic diseases, larger body size, or pregnancy. (Sen 1990, p. 116)

One of the crucial insights of the capability approach is thus that the conversion of goods and services into functionings is influenced by a range of factors, which may vary greatly from person to person. In the capability approach, a distinction is usually made between personal, social, and environmental conversion factors. The quote of Sen above mentions a couple of personal conversion factors – which are internal to the person – in relation to food resources. An example of an environmental conversion factor is climate; depending on the climate in one's living area, a certain type of house may or may not provide adequate shelter. The society in which one lives gives rise to social conversion factors, for example, the availability of nearby schools may be of no use to a girl if gender norms prevent her from taking advantage of this opportunity. In short, the fact of immense human diversity makes that a focus on capabilities is more informative of human well-being than a focus on mere resources. The main reason why capability theorists prefer these capabilities over utility or preference satisfaction is the existence of a phenomenon which Sen has called "adaptive preferences":

> Our desires and pleasure-taking abilities adjust to circumstances; especially to make life bearable in adverse situations. The utility calculus can be deeply unfair to those who are persistently deprived [. . .] The deprived people tend to come to terms with their deprivation because of the sheer necessity of survival; and they may, as a result, lack the courage to demand any radical change, and may even adjust their desires and expectations to what they unambitiously see as feasible. (Sen 1999, pp. 62–63)

Thus, if the deprived are happy with their lot in life, we cannot, according to the capability approach, conclude from this that there is no injustice in their situation.

[5] In philosophical terms, these capabilities are – at least to some degree – incommensurable.

[6] It may be that increasing someone's capability for social affiliation may turn out to be helpful as a *means* for expanding this person's capability to be well nourished – yet they are both also ends in themselves and that is where the problem of trade-offs occurs.

The capability approach thus chooses to conceptualize well-being in terms of a person's capability set and development as a process of expanding these capabilities. In this process of development, capabilities can, Sen argues, be both means and ends. For example, a person's capability to be healthy is intrinsically valuable (as an end in itself), but may also be valued instrumentally because it contributes to a person's capability to be part of a community. It should furthermore be noted here that Sen and Nussbaum use "human capabilities" as an ethical category; the term refers to those capabilities of an individual that are valuable or salient from an ethical perspective. Some capabilities may be trivial from the perspective of justice and development. Sen (1987), for example, is highly skeptical about a new brand of washing powder expanding valuable human capabilities, as advertisers tend to claim. Other capabilities may be outright undesirable to promote – Nussbaum (2000), for example, gives the example of the capability for cruelty. And a large number of more concrete capabilities will only be morally relevant because they are instrumentally important to or constitutive of the human capabilities that we ultimately or intrinsically value.

Not surprisingly one important debate within the capability approach is about which capabilities matter and who (how, when) is to decide this. This is actually one of the main topics on which Sen and Nussbaum – the former having a background in economics and the latter in philosophy – differ of opinion. Nussbaum has, after extensive discussion with people worldwide, identified a list of 10 central categories of human capabilities that are needed for living a life in conformity with human dignity, in which people can properly exercise their human agency (see "Appendix" for details of what falls under those categories):

1. Life
2. Bodily health
3. Bodily integrity
4. Senses, imagination, and thought
5. Emotions
6. Practical reason
7. Affiliation
8. Other species
9. Play
10. Control over one's environment – both political and material

She claims that justice requires bringing each and every human being up to at least a certain threshold for each of the capabilities on her list. Although Sen gives plenty of examples of important capabilities in his work, he has always refused to make such a list. His reasons are that the proper list of capabilities may depend on purpose and context and should be a result of public reasoning and democracy, not something a theorist should come up with. Democracy, public deliberation, and participation are – because of this debate about making a list of capabilities or not

and because of the value attached to human agency – also frequent topics of reflection and discussion among capability theorists (see, e.g., Crocker 2008). It is recognized by both Sen and Nussbaum that from an ethical perspective not only outcomes in terms of expanded capabilities matter but also the process through which these changes are brought about – and out of respect for people's agency, in principle participatory processes are to be preferred from a moral perspective.

Various other topics and questions also feature in the literature on the capability approach. One is, not surprisingly, the question of how to operationalize the capability approach (see, e.g., Comim et al. 2008). This includes questions on how to identify, rank, weigh, or trade off relevant capabilities in policy/project applications, on which no consensus exists. As Alkire (2005a) explains, "operationalizing is not a one-time thing," but something that is dependent upon such things as country, level of action, and the problem at hand. One of the many challenges is that it is hard to measure capabilities, as they (a) refer to the possible and not just to the realized and (b) are complex constructs depending on both an individual's internal characteristics/capacities and his/her external environment. A challenge is furthermore how to "aggregate" over people while not losing sight of the fact that a capability approach emphasizes that each and every person needs sufficient capabilities to lead a flourishing life. These questions and challenges also appear in a design application of the capability approach and will be addressed in section four.

## The Complex Relation Between Technology and Human Capabilities

The capability approach has over the past decades been applied in different ways (Robeyns 2006), such as the assessment of small-scale development projects (including projects involving the introduction of a technology; see, e.g., Fernández-Baldor et al. 2012; Vaughan 2011), theoretical and empirical analyses of policies (this may also concern technology policy or technology assessment; see, e.g., Zheng and Stahl 2012), and critiques on social norms, practices, and discourses (e.g., the ICT4D discourse; see Zheng 2009; Kleine 2011). Many of the applications so far have been concerned with backward-looking assessment and evaluation, but of course for advancing justice, well-being and development forward-looking, "prospective" applications should also receive attention (Alkire 2008), meaning that we should investigate how the expansion of human capabilities can successfully be brought about. In general terms:

> For some of these capabilities the main input will be financial resources and economic production; but for others, it can also be political practices and institutions, [. . .] political participation, social or cultural practices, social structures, social institutions, public goods, social norms, traditions and habits. (Robeyns 2005, p. 96)

Technologies could, of course, also be important inputs or means for the expansion of valuable capabilities, and indeed, increasing attention is paid to technology within the scholarly community and literature on the capability approach.[7] Before we can start to explore how the capability approach could be relevant to the design of technical artifacts, it is important to gain some basic understanding of the way in which such artifacts are related to human capabilities. As Zheng (2007) rightly noted, the capability approach – being a general normative framework - "offers little about understanding details of technology and their relationship with social processes," nor about the relations between human capabilities and technology. For this we will thus also have to turn to additional theorizing and/or empirical studies on technology.

The first thing that is important to realize is that human capabilities as discussed in the capability approach are "combined capabilities" (Nussbaum 2000), as their existence depends on a combination of two things. Only if we take both into account do we get a picture of what a person is *realistically* able to do and be in life. The first concerns internal capacities of a person, which includes both bodily and mental capacities, both innate and realized through training. The second concerns – as Nussbaum expresses it – "suitable external circumstances for their exercise," which includes the individual's embedding in institutions and practices and her/his access to resources. The latter includes technical artifacts, which are arguably – in addition to internal capacities and social structures – a third constitutive element of individual human capabilities (Oosterlaken 2011). This does not mean, of course, that technical artifacts are always effective in actually expanding valued human capabilities. As Lawson (2010) explains, "for the extension in capabilities to be realised the artefacts or devices which are used to extend the capability must be enrolled in both technical and social networks of interdependencies." It is thus the continuous interactions between these elements – the individual, technical artifacts, physical circumstances, and social structures[8] – that determine this individual's human capabilities.

A technical artifact that Sen has occasionally referred to, namely, a bicycle, may serve as an illustration. All bicycle owners are equal in terms of their possession of this resource, but people with certain disabilities will obviously not gain an increased capability to move about as a result of this bicycle (Sen 1983, 1985). One could also think of other things obstructing or facilitating the expansion of human capabilities by means of bicycles. Arguably, a person in the Netherlands – which has good roads and even many separate bicycle lanes – may gain more capabilities from owning a bicycle than a Bedouin in the desert.

[7]For example, in September 2009, the thematic group "Technology & Design" was established under the umbrella of the Human Development and Capability Association (HDCA). For a review of literature that has appeared on the topic until 2011, see the introductory chapter of the edited volume "The Capability Approach, Technology and Design" (Oosterlaken 2012).

[8]Social structures, in turn, are increasingly composed of both humans *and technical artifacts*, which is reflected in the phrase "socio-technical systems."

And if cultural norms and practices prevent women from using bicycles, as was the case in the early history of bicycle development in Europe (Bijker 1995), having a bicycle will not contribute much to capability expansion for these women either. The capability approach would acknowledge the relevance of all such contextual factors (bodily abilities, roads, supportive cultural norms) under the label "conversion factors" (already introduced in the previous section). The bicycle example may also be used to illustrate Sen's distinction between capabilities as means and ends; for some people, there may be intrinsic value in the capability to move about; a mountain biker could, for example, appreciate the sense of "flow" and freedom and the outdoor experience that the activity of cycling itself may offer. For many others, the capability to move about with a bicycle may be merely of instrumental value, as, for example, it may contribute to one's capability to visit friends (which would fall under Nussbaum's category of "affiliation") or to one's capability to exercise and in that way maintain good health. Even more indirectly, having a bicycle may contribute to one's livelihood opportunities, which could in turn again contribute in diverse ways to some of the 10 intrinsically valuable capabilities on Nussbaum's list. Of course, it is very well possible that one person values both the intrinsically valuable and the instrumental capabilities that a bicycle expands.

The example so far concerns a technical artifact that expands the capabilities of its individual users – whether direct or indirect. Yet many technologies influence our capabilities as individuals not because we *use* them, but because they are embedded in the socio-technical systems, institutions, and practices in which we are also embedded as an individual. For example, new medical technologies often lead to changes in health-care institutions and practices, and these may in turn have an impact – either positive or negative – on human capabilities. New ICTs change the ways in which governments and politicians go about their daily business, which may in turn have consequences for an individual's capability to have control over his/her political environment. Technology is also related to our culture and values in complex ways, which in turn is a relevant factor influencing people's capabilities (see, e.g., Nussbaum 2000, on culture and the capabilities of women in India). To get back to the bicycle example, Bijker (1995) concludes from his historical study of bicycle development in Europe that "the first cycles in fact reinforced the existing 'gender order,'" while "it later became an instrument for women's emancipation." Furthermore, as Coeckelbergh (2011) has pointed out, new technologies may influence our interpretation of what certain abstract capabilities, such as those on Nussbaum's list, mean. For example, ICTs such as social networking sites have arguably not merely expanded our capabilities for affiliation, but also challenged and changed our understanding of what it means to be able to engage in meaningful relations with others. Adding to the complexity is that it is conceivable that a technology expands the capability set of one category of individuals while simultaneously diminishing it for another, or influences one capability positively and another negatively, or has positive direct capability effects and negative indirect capability effects, or negative impacts on the short term and positive on the long term.

Either way, the capability approach – with its normative position that each and every person ought to have certain valuable capabilities – suggests that in the end these technologies should be evaluated in terms of their capability impacts.[9] To fully do so would require extensive empirical research, which may sometimes be – as Alkire (2010) has likewise pointed out for the relation between social arrangements and capabilities – very complex and difficult to do. The general picture that arises from the relevant literature is thus that the relation between technology and valuable human capabilities is not simple and straightforward, but dynamic and complex. Some of the implications will be addressed in section four, which discusses, among others, the epistemological challenge that designers will face in a "well-being usage" of the capability approach in design.

## The "Narrow" Application of the Capability Approach: Well-Being

As said, this chapter distinguishes between two somewhat different, although not completely separated, ways of linking the capability approach to design. In the "broad" usage the capability approach (see section "The Aggregation Challenge for a Well-Being Application") is seen as encouraging, taking a broad range of values and concerns into account in design, such as inclusiveness, agency, participation, and justice. In the *"narrow" or "well-being usage" of the capability approach* for design, the capability approach is used as a forceful reminder of the importance of human well-being and more importantly a convincing perspective on how well-being should be conceptualized and evaluated within design, namely, in terms of human capabilities. We may also call this design for capabilities.[10]

The discussion in the literature on resource possession/access versus capabilities as the best indicator of well-being is quite relevant to the design of technical artifacts. It draws the designer's attention to personal, social, and environmental "conversion factors" that should be in place before a certain artifact (merely a resource or means) can truly contribute to the expansion of valuable human capabilities (its ultimate end). In combination with the proactive design for values approach, this suggests that in order to make a meaningful contribution to improving human well-being, one should already *anticipate* these factors during the design process and try to choose design features in response to these factors. As such, the capability approach could provide an antidote to any "product fixation" that

[9]Although it is acknowledged by capability theorists that other evaluation criteria may also play a role.

[10]In a previous publication (Oosterlaken 2009), I called this "capability sensitive design, a variety on the term 'value sensitive design'" (VSD). Yet VSD is a specific approach to taking values into account in design, as developed by Friedman and colleagues. This handbook uses "design for values" for the more general idea to include values in the design process, although occasional reference to the work of Friedman and colleagues is made as well.

engineers/designers – like the economists accused of "commodity fetishism" by Sen (1985, 1984) – may suffer from on occasion.[11]

A proposal for such a "narrow" or "well-being usage" of the capability approach can be found, for example, in the work of some authors reflecting on "care robots," robots meant to contribute to the care of elderly people. Coeckelbergh (2009, 2012), a philosopher of technology, has proposed that such technologies should be evaluated in terms of their impact on the ten capability categories listed by Nussbaum. Following the proactive attitude of value sensitive design, this implies of course that we should already address these valuable human capabilities during the design phase of robot caregivers. According to Borenstein and Pearson (2010):

> . . .a typical motive for introducing robots into an environment has been to maximize profits by replacing human workers. Yet bringing robot caregivers onto the scene could also be motivated by the obligation to meet core human needs. This is a key advantage of the capabilities approach, since it should inform the design and use of robot caregivers in such a way that the 'human' in human-robot interaction is maintained. (p. 285)

More specifically, "by applying the capabilities approach as a guide to both the design and use of robot caregivers," philosophers Borenstein and Pearson say, "we hope that this will maximize opportunities to preserve or expand freedom for care recipients." The capability approach is thus used as part of an argument to put the well-being of this group of people central in design.

Another example can be found in the joint work of philosopher Colleen Murphy and civil engineer Paolo Gardoni on the capability approach and technological risks, more specifically risks related to infrastructural works. In one of their recent writings (Murphy and Gardoni 2012), they address engineering design and note that the existing

> reliability-based design codes only focus on probabilities and ignore the associated consequences [. . .] there is a need for a risk-based design that accounts in a *normative and comprehensive way* for the consequences associated to risks. (p. 174, emphasis is mine)

The capability approach is, according to them, able to fulfill this need. A "central principled advantage" is that the capability approach "puts the well-being of individuals as a central focus of the design process." The approach suggests that the negative consequences associated with risks should be expressed in terms of a range of morally salient capability deprivations. Furthermore, "a capability-based design can provide," Murphy and Gardoni claim, "some guidance to engineers as they make trade-offs between risk and meeting other design constraints, some of which may be also translated in terms of capabilities."

[11] An example may be found in Derksen (2008). She concludes that tissue engineers working on heart valves often have a limited conception of functionality and are very much focused on trying to mimic nature, while according to Derksen, they should be more concerned with the impact of the biotechnologies they develop on people's capabilities to play sports, going through pregnancy, etc. – so the sort of "beings and doings" that people have ultimately reason to value. Derksen does, by the way, not refer to the capability approach – even though what she says seems to fit in very well with that approach.

Up to present proposals for a well-being usage of the capability approach in design have, to my knowledge, not yet been followed by real-world applications. In all fairness it should, however, be acknowledged that many designers/engineers are already very aware of the importance of well-being and of taking "conversion factors" into account, even though they may not be expressing it in the same language as capability theorists. A call for structural attention for some such factors can be found, for example, in the appropriate technology movement.[12] For example, development organization practical action – which has roots in the appropriate technology movement – introduced podcasting devices in a rural area in Zimbabwe. Podcasts were recorded on topics in the area of health and cattle management (e.g., how to treat sick cows). The choice for a voice-based technology was already a response to an important personal conversion factor, namely, the illiteracy of a significant proportion of the inhabitants of the area. The exact design features were furthermore discussed taking other relevant factors into account. Important choices were that for speakers instead of headphones (in response to a common African cultural practice, e.g., sitting and sharing under a village tree) and between recharging batteries with the use of solar panels or the electricity grid (in response to local infrastructural problems). There was thus no unquestioned assumption that introducing this or that state-of-the-art ICT or technical resource could be equaled to "development" (Oosterlaken et al. 2012). Yet even if technologists/designers are already aware of the importance of conversion factors, the capability approach could still contribute by providing criteria to evaluate the importance of such factors and judge the success of such design efforts explicitly from a normative perspective, namely, in terms of people's well-being, conceived as the expansion of intrinsically valuable human capabilities.

The view of the capability approach that a range of incommensurable capabilities is needed for a good human life may also help designers to develop a broader perspective on the impact of their products on people's lives. The redesign of a silk reeling machine used in livelihood projects of an Indian development organization can illustrate this. This project was directly contributing to the women's basic capabilities to sustain themselves and their families. The new design solved problems like energy loss during reeling, failing materials, yarn quality problems, safety issues, and physical problems for the reeling women. The development organization was pleased with the new machine and took it into production. Looking back on the project years later, after being immersed in the capability approach, designer Annemarie Mink realized that she had quite uncritically accepted one part of the assignment: the machine should be made light and movable, in order to be suitable for usage at women's homes. New inquiries taught her that the reason for this design requirement was general unhappiness – mainly of the men in the villages – with the women having to work in silk reeling centers, which goes against a persistent cultural norm that women should stay home as much as possible.

[12]For a more extensive discussion of the appropriate technology movement in relation to the capability approach, see Oosterlaken et al. (2012) and Fernández-Baldor et al. (2012).

The women, however, actually appreciated being able to work in the reeling centers (Mink et al. forthcoming). How certain values and norms existing in India negatively affect the quality of life of these women, depriving them of central human capabilities, has been described impressively by Nussbaum (2000, 2011). A capability of affiliation, including being able to engage in various forms of social interaction, is on her list of 10 central human capabilities. The possibility to connect with other women in silk reeling centers could be valuable not only intrinsically but also as a means toward their further empowerment. However, in practice the freedom of women to choose to work in these reeling centers is restricted in the name of culture. And the design of the new machine turned out to facilitate this. An explicit consideration of the well-being of women in terms of Nussbaum's full list of capabilities during the design phase might have led to a different project outcome.

The capability approach, especially when illustrated with such cases from design practice, may contribute to increasing designer's sensitivity to such ethical issues. And while the capability approach provides concepts and ideas that are helpful in deliberating about them, creative value sensitive design may at least in some cases contribute to finding concrete solutions. The design for capabilities approach argued for in this section does, however, raise some challenges, so some recent work by Van de Poel (2012, unpublished draft book chapter) makes clear. He identifies two challenges for design for well-being more broadly, namely, an epistemological and an aggregation challenge. Van de Poel does discuss Nussbaum's capability list as one possible interpretation of design for well-being, yet the focus of this chapter allows for a more in-depth discussion of these challenges in relation to the capability approach.

## The Epistemological Challenge for a Well-Being Application

If a designer chooses to concentrate on the capability impacts of a product for its direct users, this raises an epistemological challenge. Van de Poel (unpublished draft book chapter) describes the challenge as follows for "design for well-being" in general:

> . . . design typically concerns products that do not yet exist; in fact design is largely an open-ended process which relates to creating a product. This means that the designers not only need knowledge of [a] what constitutes well-being for users and how that well-being might be affected by new technologies, but they must also [b] be aware that such knowledge needs to be translated into, for example, design requirements, criteria or technical parameters that can guide the design process

Let us start with sub-challenge [a]. As was explained before, Sen leaves it rather open which capabilities constitute well-being, while Nussbaum's version of the capability approach provides more guidance. However, a feature of Nussbaum's list of 10 intrinsically valuable capabilities is its "multiple realizability" (Nussbaum 2000, p. 105). It thus still needs to be investigated what these rather abstract

capabilities, such as the capability for play or affiliation, could – with preservation of their moral import[13] – mean exactly in the context or culture for which the design is meant. Moreover, the effect of new technologies on human capabilities, so I argued in section two, is dynamic and complex. It may be good for designers to be aware that this is the case. Yet for both practical and epistemic reasons, it does not seem realistic to expect them to anticipate and/or influence *all* capability effects of the artifacts that they help create. Their technical and empirical investigations, as part of a design for capabilities process, will need to be focused on the capabilities, conversion factors, and issues that seem most salient and relevant to the design challenge in question.

An obvious and often defensible curtailment will be to concentrate on the well-being of the expected direct users of a technology. One can doubt, says Van de Poel (2012), "whether there is a moral imperative for designers to *increase* the well-being of other stakeholders besides users." In contrast, the moral imperative not to harm other stakeholders cannot be dismissed that easily, which may sometimes mean that attention needs to be paid to the capability impacts for nonusers. Take the example provided by Murphy and Gardoni: infrastructural works may also come with risks for nonusers, which may be conceptualized as diminishing the security of their capabilities. There are strong ethical reasons for designers to take this possible harm into account.[14] In any case, an extensive discussion of the moral obligations of designers is beyond the scope of this chapter. The point here is that, as part of design for capabilities, there is a need for integrated conceptual and empirical investigations[15] addressing the relevance and meaning of certain capabilities and the contribution that a certain technology/design could make to expand those. Design for capabilities requires, as Van de Poel (2012) remarks for design for well-being in general, "more than just the identification of user demands by means of surveys or marketing research." One thing that may be beneficial for design for capabilities is more ethnographic style research for better understanding of the relation between technology and human capabilities in light of the local context and good life views.[16]

In another article, Van de Poel (forthcoming) has reflected on sub-challenge [b], translating values into design requirements, criteria, and so on. This process, so he warns, "may be long lasting and cumbersome"; it "may require specific expertise,

[13]What is meant by the latter is that a certain more concrete conceptualization of an abstract capability should do justice to or at least cohere with the reasons we have to consider the abstract capability to be valuable in the first place.

[14]The distinction made here mirrors the distinction made by philosophers between positive duties of benevolence and negative duties not to harm, where the latter is in general considered to be stronger and less controversial than the former. But Van de Poel notices that "increasing or maximizing user well-being is often mentioned or assumed as goal in design."

[15]Reference is made here to the "tripartite methodology as proposed by Friedman et al. (2001), consisting of integrated empirical, technical, and conceptual investigations. See also the entry on value sensitive design elsewhere in this handbook.

[16]I take this suggestion from an article by Ratan and Bailur on the capability approach and "ICT for Development" (2007).

sometimes from outside engineering"; it "is value laden," "can be done in different ways," and is "context-dependent." That last point may be considered to be especially important from the perspective of the capability approach, considering its emphasis on human diversity and the great variety of personal, social, and environmental conversion factors. A central idea in Van de Poel's paper on how to translate values into design requirements is that of a "value hierarchy" going from abstract values, via norms to concrete design requirements – where each of these three main layers may have sub-layers again. An example that he gives is that of animal welfare as a central value in the design of chicken husbandry systems. This value may be translated into norms such as "presence of laying nests," "enough living space," and so on. The latter norm could in turn be translated in a requirement to have at least 1,100 $cm^2$ usable area per hen.[17] According to Van de Poel:

> The reconstruction of a values hierarchy makes the translation of values into design requirements not only more systematic, it makes the value judgments involved also explicit, debatable and transparent. (Van de Poel forthcoming)

The reconstruction of value hierarchies can be helpful, even though – as Van de Poel notices – merely describing a value hierarchy does not directly solve possible disagreements about such translations.

This idea of a value hierarchy can, it seems to me, also be put to use in the context of design for capabilities, helping designers to address the epistemological challenge. One of Nussbaum's 10 capabilities – or a context-dependent interpretation of it – could be put at the very top of the value hierarchy of a design for capabilities project. In the layer below one could put – among others – more concrete capabilities, which are important *for the sake of* the high-level capability. "For the sake of," Van de Poel explains, can be "seen as the placeholder for a number of more specific relations." A certain capability could, for example, be either a *constitutive part* of a higher-level capability or be *a means toward* that capability. Let me give some examples. One's capability to be free of malaria could be said to be *constitutive of* one's capability for bodily health – to which designers may, for example, contribute by creating a new malaria diagnostic device that is suitable for usage in rural areas in developing countries. As we have seen, many conversion factors may stand in the way of such a device leading to the expansion of the capability in question for, say, villagers in India. These factors can be an important source for norms and subsequent concrete design requirements – for example, the fact that local health-care workers have little education may lead to a norm that the device should have a simple and intuitively clear user interface.[18]

[17]Van de Poel (forthcoming) points out that "the relation between the different layers of a values hierarchy is not deductive. Elements at the lower levels cannot be logically deduced from higher level elements. One reason for this is that the lower levels are more concrete or specific and that formulating them requires taking into account the specific context or design project for which the values hierarchy is constructed."

[18]This example is inspired by an actual design project described in Kandachar et al. (2007).

Or take the example of a project to design a walker for elderly people. One's capability to move around can be seen as an end in itself, but it can also be considered as *a means for* one's capability for affiliation. In the latter case one can argue that one of the norms should be that one can also comfortably use the walker as a temporary seat when encountering people in the street that one would like to talk to. What this example nicely illustrates is that a technical artifact may often – direct or indirect, intended or unintended, or positively or negatively – affect a range of different capabilities, even though the primary function of a technical artifact seems closely tied to one specific capability. Nussbaum's list of capabilities needed for a flourishing human life may invite designer to always investigate the potential impact of a project on a range of capabilities, instead of a single one. In both examples – the malaria diagnostic device and the walker – the norms identified still need to be further translated into concrete design requirements, which will make sure that the interface will be clear enough, respectively, and the seat comfortable enough.

## The Aggregation Challenge for a Well-Being Application

In addition to the epistemological challenge, Van de Poel (2012) rightly notices that design for well-being will run into an aggregation problem, which

> . . . arises due to the fact that a design does not affect the well-being of just one person, but rather that of a range of people. This raises the question of how the well-being of these people should be aggregated so that it can be taken into account in the design process. If one believes that well-being constitutes plural and incommensurable prudential values, as some philosophers [. . .] have suggested, then an aggregation problem arises with respect to how these values can, or cannot, be aggregated into an overall measure of well-being. (p. 296)

As was explained in section one, the capability approach in general also faces both these problems of aggregation over (a) a range of people while not losing sight of the moral worth of each and every individual and (b) plural, incommensurable capabilities (see, e.g., Comim 2008). The incommensurability of values, Van de Poel (unpublished draft book chapter) notes, "limits the applicability of [maximizing] methods such as cost benefit analysis and multi-criteria analysis which are often used in technical design to choose between different conceptual design solutions." Luckily, he says, there exist alternative methods "not unfamiliar to the field of design." He distinguishes between two different situations. The first is where design is supposed to contribute to elementary capabilities in contexts of great poverty. Here the solution that Van de Poel proposes is – in line with Nussbaum's position – to "set thresholds for all the relevant capabilities and to look for a design that reaches all of these thresholds."

The second situation is contexts of more welfare where "one aims to find a design that contributes to the overall well-being of users." Here the focus will be on more intricate and complex capabilities rather than basic capabilities. The solution that van de Poel proposes consists of several elements. A basic step is to "select a user group that shares a comprehensive [life] goal and/or a vision of the good life,"

a step "which avoids the need to aggregate the well-being of people who have different, incompatible" goals or visions. The idea is then to come up with a mix of specific values (or capabilities, in the context of this chapter) to which a technology may contribute and then to design a product "that enables this mix as much as possible." Van de Poel (2012) hastens to add here that this

> does not imply a maximising approach to well-being. The focus is on the mix of values [or capabilities] rather than on maximising an overall measure of well-being. The focus is also not on maximising each of the prudential values [or capabilities] in isolation, because it is usually the mix of values [or capabilities] that contributes to the overall goal rather than the values [or capabilities] in isolation. (p. 303)

Incommensurability of capabilities thus need not become a problem if creative design solutions enable us to expand all of them rather than to make a trade-off between them.[19]

Van de Poel's idea of focusing on a mix of capabilities rather than on single ones shows some resemblance to the idea of a "capability innovation" that was introduced by Ziegler (2010). Building on Schumpeterian economics, which views development as a process of economic innovation in the sense of "new combinations in terms of new goods, new methods of production, and so on," Ziegler defines *social* innovation as "the carrying out of new combinations of capabilities." Ziegler views – in line with the capability approach – capabilities as both ends in itself and means toward other capabilities, emphasizing that the "relations between the capabilities" are "especially important" in his concept of capability innovation. Of course new products and their design details may be an essential element in the success of "capability innovations," as a case study in a later paper of the same author makes clear (Ziegler et al. forthcoming).

A design case discussed by Oosterlaken (2009) may be taken to illustrate the idea of "capability innovations." It concerns a project on tricycles for disabled people in Ghana, executed by industrial design engineering students (Kandachar et al. 2007). Both the local context and entrepreneurial opportunities were carefully taken into consideration. During exploratory field studies, it was discovered, for example, that "the major part of the disabled population is willing to work but cannot find employment" and that "the Ghanaian society is annoyed by disabled who are begging on the street." The newly designed tricycle has a cooler in front so that disabled users are able to make a living as street vendors selling ice cream and other frozen products. To make this a sustainable development success, it was investigated how to embed this artifact in a larger plan and network also involving a local metal workshop being able to produce and repair the tricycles and a supplier of products to be sold.[20] It can be considered a capability innovation in Ziegler's sense, as it involves a clever combination of simultaneously expanding for these

[19]Van den Hoven et al. (2012) extensively argue along these lines concerning incommensurable values and moral dilemmas more broadly.

[20]A pilot was subsequently executed.

disabled the capabilities for mobility, earning a living (and hence basic capabilities related to survival and health), social participation, and self-esteem.

## A "Broad" Application of the Capability Approach: Agency

Having discussed the narrow or well-being usage of the capability approach for design in some detail, I would now like to move to a "*broad" usage of the capability approach* in the context of design. In a broader usage, the capability approach not only is seen as highlighting the importance of individual well-being and conceptualizing this in terms of human capabilities but is also seen as taking aboard a wider range of values, most importantly agency and justice. The importance of agency in the capability approach is, as was already explained, among others, reflected in the approach's defense of capabilities instead of functionings as a policy goal. The idea behind making a distinction between capabilities and functionings is – as explained before – to be respectful of people's agency and their views on the good life by focusing on expanding their capabilities without forcing them to realize the corresponding functionings. Capability scholars acknowledge, however, that there are sometimes reasons why a focus on functionings instead of capabilities may be justified for evaluative purposes (see, e.g., Robeyns 2005, p. 101).

In the case of the design of technical artifacts, one might also wonder if it is sensible to uphold this distinction; is an artifact that does not lead to an increase in the functioning(s) that the designer aimed at not a failure? This depends. If people lacking the functioning have freely chosen not to realize it, we generally need to accept and respect this. But if the functioning in question is absent on a massive scale, this may warrant further investigation. Has the designer failed to grasp what capabilities are important to people's lives and has therefore nobody chosen to use the artifact to realize the corresponding functionings? This would be a matter of people exercising their agency. Or are there perhaps disruptive conversion factors in play that nobody foresaw and has the design therefore not really enabled people to realize these functionings? This would mean that the design has not lead to empowerment or an increase of agency. These two causes, which can be distinguished when looking through the lens of the capability approach, obviously ask for different responses.

The capability-functioning distinction may also make designers aware of how much choice they are giving the users (see, e.g., Steen et al. 2011; Kleine et al. 2012). Are the products of design merely expanding people's capabilities, or are they somehow forcing people into certain functionings? The capability approach can, for example, be seen to provide a critical perspective on the so-called behavior steering technology, even when designed to contribute to well-being – as it will mean pushing people into certain functionings. This might indicate insufficient respect for people's own agency, although Nussbaum (2000) has argued that "we may feel that some of the capabilities [like that of being healthy] are so important, so crucial to the development or maintenance of all others, that we are

sometimes justified in promoting functioning rather than simply capability, within limits set by an appropriate concern for liberty." The concept of "adaptive preferences," prominently present in the capability approach literature, also implies that respect for people's agency should not be taken to mean that designers always need to respect each and every preference that people happen to have. This is an important point, as preference satisfaction – or something akin, like desire satisfaction or happiness – is what design often aims at (Oosterlaken 2009; Van de Poel 2012). The capability approach offers a richer, less subjective understanding of human well-being, which may challenge designers to develop a critical and deliberative attitude and look beyond what people superficially seem to want. However, too easily labeling someone's preferences as "adaptive" would lead to unjustified paternalism – which especially in the intercultural context of "design for development" may quickly become an issue.[21] It is hard if not impossible to provide general guidelines on how to balance these different concerns, abstracted from the details of concrete cases. In short, the capability approach does not offer quick and easy guidelines for designers, but rather a conceptual framework that helps highlighting and discussing important issues.

Furthermore, capability theorists connect agency not only to outcomes in terms of the expansion of human capabilities but also to the process leading to these outcomes. In the capability approach people are not viewed as passive patients to be helped, but as agents in charge of their own development process. Hence, the literature on the capability approach pays attention to participatory processes and democratic deliberation as both an expression of people's agency and a way to expand their agency (see, e.g., Crocker 2008). A connection can be made here with the so-called participatory design. According to Nieusma (2004), this "has developed into a well-articulated, well-justified methodology for user participation in design processes" and should be all about "coping with disagreements." He regrets, however, that "increasingly, participatory design methodologies are used to advance the goals of user-centred design without emphasizing the inclusion of marginalized perspectives in design processes." According to Buchanan (2001) as well, designers often "reduce [their] considerations of human-centred design [which often involves users in the design process] to matters of sheer usability."

The capability approach may be helpful in revitalizing the ideals of participatory design (Oosterlaken 2009). A parallel can be drawn here with participatory methods in development cooperation. In practice, says Frediani (unknown date), these methods often do not meet the expectations, being "sometimes used merely as a tool for achieving pre-set objectives" and not as a process for true empowerment and improvement of people's lives. He argues that "participatory methods need to be complemented by a theory that explores the nature of people's lives and the relations between the many dimensions of well-being." This theory, he says, should be comprehensive, but flexible and able to capture complex linkages between

[21]See, e.g., the blog of Bruce Nussbaum titled "Is Humanitarian Design the New Imperialism?" (http://www.fastcodesign.com, blog from July 7th 2010).

(aspects of) poverty, intervention, participation, and empowerment. He feels that the capability approach is able to offer exactly that. Similarly, Frediani and Boano (2012), who focus on urban design, note "a surprising lack of literature investigating the conceptual underpinnings of participatory design and its implications in terms of practice," a gap which – according to them – could be filled with the help of the capability approach.

## A "Broad" Application of the Capability Approach: Justice

Going a step beyond "mere" participation in a process where professional designers are still in the lead is proposed by Dong (2008), who believes "that the capabilities approach offers one avenue to situate design practice as part of an endeavour of social justice." His focus is on the design of civic works and the built environment. He argues that such design is intimately connected to people's health and identity and therefore Dong proposes to "add 'control over the design and production of civic building' to Nussbaum's list as sitting astride political and material control." However, one could easily extend Dong's argument to the design of technical artifacts more broadly; if we combine the fact that these are nowadays ubiquitous in all domains of human life with insights on the "politics" (Winner 1980) and "value ladenness" (see, e.g., Radder 2009) of such artifacts, it seems that Nussbaum's description of what control over one's environment entails (see "Appendix") is too narrow and should include control over one's *designed* surroundings (including buildings and other artifacts). But back to Dong's (2008) line of argument:

> Public policies can effectively remove public engagement in the name of expediency. [. . .] Thus, what the urban poor in developing countries and citizens in developed countries share is the problem of enacting a policy of design that reflects the values of the people. [. . .] People have the right to user participation in design only if there are effective policies to make people truly capable of design. So what is needed is not user participation in design as a counterforce to the power of designers [. . .] but instead a design culture of pluralism with effective means for achieving it. (p. 77)

Dong argues that from a justice perspective, we should pay attention to citizens' capabilities to design *themselves* and in this way enable them to co-shape their life world. For this purpose he fleshes out a set of instrumentally important capabilities that citizens would need to do design, which could become object of (inter)national design policy. The categories that he distinguishes are information, knowledge, abstraction, evaluation, participation, and authority. Dong points out, in line with the capability approach, that "asymmetries in capability to *do* design may arise from differences between people and socio-political barriers" and that design policy should thus address both these internal and external factors. As Nichols and Dong (2012) explicate: gaining design capacity or skill – as the "humanitarian design community" apparently promotes – is not enough for truly gaining the "capability to design." The latter may, for example, be inhibited by political factors even though design skills are present.

Not only the capability approach and participatory design could be fruitfully connected but also inclusive/universal design and the capability approach. What the latter two share (Oosterlaken 2012) is an awareness of the pervasiveness and importance of human diversity and the injustice to which neglecting this may lead. The paradigm example here is buildings being inaccessible for wheelchair users – in the language of the capability approach, we could say that personal conversion factors in this case hamper the conversion of resources into valuable capabilities. Nussbaum (2006, p. 167) considers such a design as a serious matter of injustice. As was mentioned in section one, injustice occurs according to her when people fall below a certain threshold level of capabilities, which may occur when human diversity is not taken sufficiently into account or in other words when conversion factors for different categories of people have not been sufficiently considered by designers.

The inclusive/universal design movement has also addressed this case of wheelchair-unfriendly buildings (Connell and Sanford 1999), by advocating designs that are usable by a wide variety of users, including but not limited to people with disabilities. Although wheelchair-friendly buildings may have become the standard by now, in other domains of design and for other user groups, the inclusive/universal design movement may still have work to do. Toboso (2011), for example, claims that there is not enough attention for diversity in the design of ICTs. He uses the capability approach to rethink disability and proposes to enrich it with the concept of "functional diversity" to support the shift in design practice that he proposes. The capability approach could learn a lot from how the inclusive/universal design has come up with solutions for the challenge of human diversity, thus contributing to the expansion of human capabilities and the practical realization of the normative ideals of the capability approach.

The universal/inclusive design movement, on the other hand, might benefit from a better acquaintance with the capability approach and the conceptual framework it provides (Oosterlaken 2012). It may help designers to get a better understanding of the ultimate aims of design and may make it possible for them to make a quite natural connection between their work and wider normative debates about justice and development. Furthermore, the degree to which a design contributes to the actual realization of human capabilities of different categories of users could be used as a yardstick to determine whether or not universal/inclusive design has achieved its moral objective.

## Looking Ahead: Some Further Challenges

As was already mentioned before, many designers are not oblivious to the considerations that a capability approach of design would highlight; in fact they regularly already take these into account, without using the capability approach's vocabulary. Yet using the capability approach could make these design considerations more explicit and therefore more open to scrutiny and debate. The capability approach has the potential – to borrow some words of Zheng (2007) – to "surface a set of key

concerns [most importantly justice, well-being and agency] systematically and coherently, on an explicit philosophical foundation."[22] The previous sections have hinted at possible benefits of more explicitly applying the capability approach to design, but of course the proof of the pudding is in the eating and that is where it is still lacking. Theorizing on the capability approach and design has only just begun, and practical experience with it is still extremely limited. One of the challenges is – so my interactions with some designers have indicated – that the conceptual framework of the capability approach is not so intuitively obvious[23] and it takes some effort to learn it. And although some designers may be motivated to plow through the many insightful books and articles of Sen, Nussbaum, and other capability theorists, it is not realistic to expect this from all designers.

One possible solution – one that I would personally expect to appeal to practical people like designers – is to develop checklists and tools based on the capability approach that designers could use in different phases of the design process. So far, these do not exist.[24] For other members of the "design for values" family, such as design for sustainability, a lot of progress has been made on this path. This comes, however, with a risk of an uncritical usage and an unhelpful simplification of the issues and dilemmas at hand. For example, various software packages exist that help designers to make a qualitative life cycle analysis of their product. In response the chapter on sustainability in a main textbook for teaching ethics to engineers (Brumsen 2011) warns engineers that these programs may create an unjustified air of simplicity. They weigh and add different environmental aspects into one final number. Thus, the software's outcomes are based on the normative considerations of the programmers, a specific way of aggregating, which may subsequently not become a topic of discussion among designers. Even more qualitatively oriented lifecycle approaches, so the author points out, still have the disadvantage of focusing on environmental impact, while leaving other aspects of sustainability, such as intergenerational justice, unaddressed.

[22]Zheng (2007) is speaking about applying the capability approach to the area of "ICT for Development (ICT4D)," and parallels may be drawn with applying it to design. She notes that "many of the issues unveiled by applying the capability approach are not new to e-development research." Yet, she feels that the capability approach is "able to surface a set of key concerns systematically and coherently, on an explicit philosophical foundation," and, "as a conceptual basis, could accommodate other theoretical perspectives in e-development," like discourse analysis, institutional theory, social inclusion, the participative approach, local adaptation, and information culture.

[23]For example, what is the difference between the function of an artifact and the concept of "functionings" in the capability approach? What distinguishes a capacity or skill from a capability?

[24]Nussbaum's list of 10 central capabilities may serve as a starting point for designers, but it has not been tested yet if and how it helps designers in their deliberations about their design project. Moreover, as discussed in section four, the list is quite abstract and applying it in design would still require quite a lot of additional work, so that "just" giving this list to designers is probably not enough.

One might say that the idea of sustainability and the capability approach share the problem of multidimensionality and incommensurability, which provides a challenge for their "operationalization" for designers. Providing designers with concrete tools in which the thinking has already been done for them does not seem the way to go for an approach that emphasizes the pervasiveness of human diversity, both in people's circumstances and characteristics and their ideas of the good life. Yet there is surely a lot of middle ground between that path and giving designers a pile of philosophical books. One could think of an inspirational portfolio of design cases analyzed with the capability approach and illustrating dilemmas encountered, in combination with exercises developed to "sensitize" designers to different ideas highlighted by the capability approach. Approaches such as Van de Poel's usage of a value hierarchy in translating values into design requirements could be further investigated in relation to the capability approach, as could other tools and approaches developed within value sensitive design more broadly. And of course there is a lot to be gained from looking at the work already done in design movements which share some ideals and insights with the capability approach, such as participatory design and inclusive/universal design.

A completely different type of challenges arises from the fact that, as was explain in section one, human capabilities concern what a person is able to do and be *all things considered*. This, so section two explained, also implies that human capabilities do not merely depend on technical artifacts and other products of design, but also on their embedding in broader socio-technical networks. This insight may result in some skeptical doubts about the possibility of design for capabilities. It seems undeniable that there are substantive limits – including epistemological ones – to the degree to which designers can take responsibility for the wider socio-technical environment in which their products will be embedded and thus for the effective creation of valuable human capabilities. This seems to be even more the case when we take into account the long-term and systemic effects of the introduction of new technologies, which may have an indirect effect on a range of valuable capabilities. One may therefore wonder if design for capabilities is not just a very nice idea that is very difficult, if not impossible, to put into practice. However, that the details of design often matter to some degree for the capabilities that technical artifacts do or do not expand seems to me just as undeniable as the limitations to the influence of design. To what degree, so we can learn from the empirical turn in philosophy of technology (Kroes and Meijers 2000), is not something that we can resolve in the abstract, for technology in general. This therefore requires further study in real-world cases.

I think though that this skeptical response raises a further issue, namely, about how we understand, organize, and practice design. Instead of strongly contrasting them or seeing them as complementary, we should perhaps rather think about merging them by thoroughly rethinking design itself and expanding its scope. The need to do so and take a "system view of design" is perhaps most salient in the context of developing countries, where even basic socio-technical networks

and infrastructures are lacking (Sklar and Madsen 2010), but even in the North this may sometimes be needed – an obvious example probably being electric cars, the introduction of which requires an integrated approach of both product design and socio-technical system development. Design for capabilities, in order to become reality, may thus need to connect to current discourses on systems and design, like perhaps "whole system design" (e.g., Blizzard and Klotz 2012) or the design of PSS or product/service systems (e.g., Morelli 2002) – however, for both, the availability of tools, methods, and design principles is still one of the challenges. Also against the background of VSD or "value sensitive design," it has been noted (Nathan et al. 2008) that "a scarcity of methods exists to support long-term, emergent, systemic thinking in interactive design practice, technology development and system deployment." Philosophers of technology may have a contribution to make to such a system-oriented endeavor to give the idea of design for capabilities more substantive content (see, e.g., Kroes et al. 2006; Krohs 2008).

## Conclusion

That the capability approach can be brought to bear on technology's design should be clear by now. Starting with an intuition that technical artifacts have in essence something to do with enabling human action, with expanding what persons are able to do and be, this chapter has explored the relevance of the capability approach – being a philosophical framework that attaches central *moral importance* to human capabilities – for the value sensitive design of such artifacts. A distinction was made between a "narrow" or "well-being usage" of the capability approach and a "broad usage" in which the capability approach is also seen as a source of insight and inspiration with respect to a wider range of values, most prominently agency and justice. Each of these three values can be conceptualized and understood with reference to human capabilities, and relations between these values have emerged in this chapter. Participatory design can be connected to both agency and justice, inclusive or universal design to both justice and well-being, and design for capabilities to both well-being and agency. And each should be done with an awareness of the broader socio-technical embedding of the object of design (see Fig. 1). In reality, however, both usages can and should often go hand in hand.

In previous work on the capability approach, the impression was given that a well-being usage of the capability approach is exclusively connected to a product-oriented application of the capability approach (Oosterlaken 2009), whereas a process-oriented application of the capability approach connects to agency (Dong 2008). Frediani and Boano (2012) reject such an "unhelpful dichotomy":

> . . .the analysis should not merely engage with the process of design, but also with its outcomes. The reason is that citizens' design freedom is shaped not merely by their choices, abilities and opportunities to engage in the process of design, but also by the degree to which the outcomes being produced are supportive of human flourishing. (p. 210).

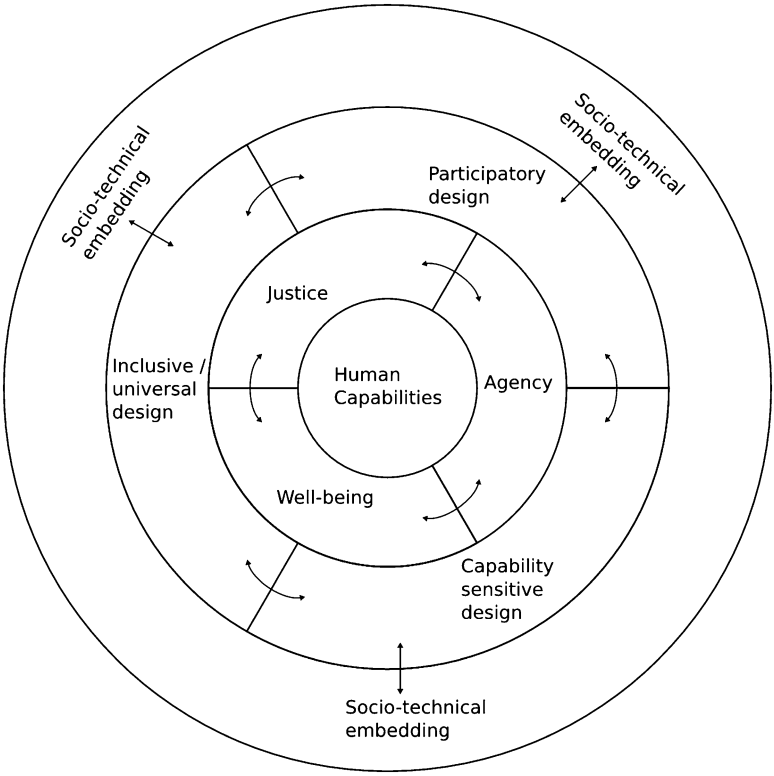



**Fig. 1** Values central in the capability approach and their relation to design

This chapter has shown that a capability approach of design should indeed embrace all these different values and design approaches.

**Acknowledgments** This research has been made possible by a grant from NWO, the Netherlands Organization for Scientific Research. The author would like to thank a number of people for their valuable feedback on earlier versions of this chapter: Annemarie Mink, Ibo van de Poel, Sabine Roeser, and Rafael Ziegler.

## Cross-References

► Design for the Value of Inclusiveness
► Design for the Values of Democracy and Justice
► Design for the Value of Presence
► Design for the Value of Human Well-Being
► Value Sensitive Design: Applications, Adaptations, and Critiques

## Appendix

### Nussbaum's 10 Central Capabilities

The complete and detailed list of central human capabilities according to Nussbaum (2002):

1. *Life*. Being able to live to the end of a human life of normal length; not dying prematurely, or before one's life is so reduced as to be not worth living.
2. *Bodily Health*. Being able to have good health, including reproductive health; to be adequately nourished; to have adequate shelter.
3. *Bodily Integrity*. Being able to move freely from place to place; to be secure against violent assault, including sexual assault and domestic violence; having opportunities for sexual satisfaction and for choice in matters of reproduction.
4. *Senses, Imagination, and Thought*. Being able to use the senses, to imagine, think, and reason – and to do these things in a "truly human" way, a way informed and cultivated by an adequate education, including, but by no means limited to, literacy and basic mathematical and scientific training. Being able to use imagination and thought in connection with experiencing and producing works and events of one's own choice, religious, literary, musical, and so forth. Being able to use one's mind in ways protected by guarantees of freedom of expression with respect to both political and artistic speech, and freedom of religious exercise. Being able to have pleasurable experiences and to avoid non-beneficial pain.
5. *Emotions*. Being able to have attachments to things and people outside ourselves; to love those who love and care for us, to grieve at their absence; in general, to love, to grieve, to experience longing, gratitude, and justified anger. Not having one's emotional development blighted by fear and anxiety.
6. *Practical Reason*. Being able to form a conception of the good and to engage in critical reflection about the planning of one's life. (This entails protection for the liberty of conscience and religious observance.)
7. *Affiliation.*
   (a) Being able to live with and toward others, to recognize and show concern for other humans, to engage in various forms of social interaction; to be able to imagine the situation of another. (Protecting this capability means protecting institutions that constitute and nourish such forms of affiliation, and also protecting the freedom of assembly and political speech.)
   (b) Having the social bases of self-respect and non-humiliation; being able to be treated as a dignified being whose worth is equal to that of others. This entails provisions of non-discrimination on the basis of race, sex, sexual orientation, ethnicity, caste, religion, national origin and species.
8. *Other Species*. Being able to live with concern for and in relation to animals, plants, and the world of nature.
9. *Play*. Being able to laugh, to play, to enjoy recreational activities.
10. *Control over one's Environment.*

(a) Political. Being able to participate effectively in political choices that govern one's life; having the right of political participation, protections of free speech and association.
(b) Material. Being able to hold property (both land and movable goods), and having property rights on an equal basis with others; having the right to seek employment on an equal basis with others; having the freedom from unwarranted search and seizure. In work, being able to work as a human, exercising practical reason and entering into meaningful relationships of mutual recognition with other workers.

## References

Alkire S (2005) Why the capability approach? J Hum Dev 6(1):115–133
Alkire S (2008) Using the capability approach: prospective and evaluative analyses. In: Comim F, Qizilbash M, Alkire S (eds) The capability approach: concepts, measures and applications. Cambridge University Press, Cambridge
Alkire S (2010) Instrumental freedoms and human capabilities. In: Esquith SL, Gifford F (eds) Capabilities, power, and institutions: toward a more critical development ethics. The Pennsylvania State University Press, University Park
Bijker WE (1995) Of bicycles, bakelites, and bulbs: toward a theory of sociotechnical change. The MIT Press, Cambridge, MA
Blizzard JL, Klotz LE (2012) A framework for sustainable whole systems design. Des Stud 33 (2012):456–479
Borenstein J, Pearson Y (2010) Robot caregivers: harbingers of expanded freedom for all? Ethics Info Technol 12(4):277–288
Brumsen M (2011) Sustainability, ethics and technology. In: Van de Poel I, Royakkers L (eds) Ethics, technology and engineering. Wiley-Blackwell, Malden/Oxford
Buchanan R (2001) Human dignity and human rights: thoughts on the principles of human-centered design. Des Issues 17(3):35–39
Coeckelbergh M (2009) Health care, capabilities and AI assistive technologies. Ethic Theory Moral Pract 13(2):181–190
Coeckelbergh M (2011) Human development or human enhancement? A methodological reflection on capabilities and the evaluation of information technologies. Ethics Info Technol 13 (2):81–92
Coeckelbergh M (2012) "How I learned to love the robot": capabilities, information technologies, and elderly care. In: Oosterlaken I, Van den Hoven J (eds) The capability approach, technology & design. Springer, Dordrecht
Comim F (2008) Measuring capabilities. In: Comim F, Qizilbash M, Alkire S (eds) The capability approach: concepts, measures and applications. Cambridge University Press, Cambridge
Comim F, Qizilbash M, Alkire S (2008) The capability approach: concepts, measures and applications. Cambridge University Press, Cambridge
Connell BR, Sanford JA (1999) Research implications of universal design. In: Steinfeld E, Danford GS (eds) Enabling environments: measuring the impact of environment on disability and rehabilitation. Kluwer, New York
Crocker DA (2008) Ethics of global development: agency, capability, and deliberative democracy. Cambridge University Press, Cambridge
den Hoven V, Jeroen G-JL, Van de Poel I (2012) Engineering and the problem of moral overload. Sci Eng Ethics 18(1):143–155
Dong A (2008) The policy of design: a capabilities approach. Des Issues 24(4):76–87

Fernández-Baldor Á, Hueso A, Boni A (2012) From Individuality to collectivity: the challenges for technology-oriented development projects. In: Oosterlaken I, Van den Hoven J (eds) The capability approach, technology and design. Springer, Dordrecht

Frediani AA (unknown date) Participatory methods and the capability approach. In: Briefing notes: Human Development and Capability Association. http://www.capabilityapproach.com/pubs/Briefing_on_PM_and_CA2.pdf. Accessed 13 June 2008

Frediani AA, Camillo B (2012) Processes for just products: the capability space of participatory design. In: Oosterlaken I, Van den Hoven J (eds) The capability approach, technology and design. Springer, Dordrecht

Friedman B, Kahn PH, Borning A (2001) Value sensitive design: theory and methods. UW CSE technical report

Gigler B-S (2008) Enacting and interpreting technology – from usage to well-being: experiences of indigenous peoples with ICTs. In: Van Slyke C (ed) Information communication technologies: concepts, methodologies, tools, and applications. IGI Global, Hershey

Illies C, Meijers A (forthcoming) Artefacts, agency and action schemes. In: Kroes P, Verbeek P-P (eds) Technical artefacts and moral agency. Springer, Dordrecht

Kandachar P, Diehl JC, van Leeuwen G, Daalhuizen J (eds) (2007) Design of products and services for the base of the pyramid, vol 2, IDE graduation projects. Delft University of Technology, Faculty of Industrial Design Engineering, Delft

Kleine D (2011) The capability approach and the 'medium of choice': steps towards conceptualising information and communication technologies for development. Ethics Info Technol 13(2):119–130

Kleine D, Light A, Montero M-J (2012) Signifiers of the life we value? – considering human development, technologies and Fair Trade from the perspective of the capabilities approach. Info Technol Dev 18(1):42–60

Kroes P, Meijers A (eds) (2000) The empirical turn in the philosophy of technology, vol 20. JAI/Elsevier, Amsterdam

Kroes P, Franssen M, Van de Poel I, Ottens M (2006) Treating socio-technical systems as engineering systems: some conceptual problems. Syst Res Behav Sci 23(2006):803–814

Krohs U (2008) Co-designing social systems by designing technical artifacts. In: Vermaas PE, Kroes P, Light A, Moore SA (eds) Philosophy and design – from engineering to architecture. Springer, Dordrecht, pp 233–245

Lawson C (2010) Technology and the extension of human capabilities. J Theory Soc Behav 40 (2):207–223

Mink A, Parmar VS, Kandachar PV (forthcoming) Responsible design and product innovation from a capability perspective. In: Van den Hoven J et al (eds) Responsible innovation. Innovative solutions for global issues, vol 1. Springer, Dordrecht

Morelli N (2002) Designing product/service systems: a methodological exploration. Des Issues 18 (3):3–17

Murphy C, Gardoni P (2012) Design, risk and capabilities. In: Oosterlaken I, Van den Hoven J (eds) The capability approach, technology and design. Springer, Dordrecht

Nathan LP, Friedman B et al (2008) Envisioning systemic effects on persons and society throughout interactive system design. In: Proceedings of DIS 2008. ACM Press, New York, pp 1–10

Nichols C, Dong A (2012) Re-conceptualizing design through the capability approach. In: Oosterlaken I, Van den Hoven J (eds) The capability approach, technology & design. Springer, Dordrecht

Nieusma D (2004) Alternative design scholarship: working towards appropriate design. Des Issues 20(3):13–24

Nussbaum MC (2000) Women and human development: the capability approach. Cambridge University Press, New York

Nussbaum MC (2006) Frontiers of justice: disability, nationality, species membership. The Belknap Press of Harvard University Press, Cambridge

Nussbaum MC (2011) Creating capabilities: the human development approach. The Belknap Press of Harvard University Press, Cambridge
Oosterlaken I (2009) Design for development: a capability approach. Des Issues 25(4):91–102
Oosterlaken I (2011) Inserting technology in the relational ontology of sen's capability approach. J Hum Dev Capab 12(3):425–432
Oosterlaken I (2012a) The capability approach and technology: taking stock and looking ahead. In: Oosterlaken I, Van den Hoven J (eds) The capability approach, technology and design. Springer, Dordrecht
Oosterlaken I (2012b) Inappropriate artefact, unjust design? – human diversity as a key concern in the capability approach and inclusive design. In: Oosterlaken I, Van den Hoven J (eds) The capabililty approach, technology and design. Springer, Dordrecht
Oosterlaken I, Grimshaw DJ, Janssen P (2012) Marrying the capability approach, appropriate technology and STS: the case of podcasting devices in Zimbabwe. In: Oosterlaken I, Van den Hoven J (eds) The capability approach, technology and design. Springer, Dordrecht
Radder H (2009) Why technologies are inherently normative. In: Meijers A (ed) Handbook of the philosophy of technology and engineering sciences. Reed Elsevier, Amsterdam
Ratan AL, Bailur S (2007) Welfare, agency and "ICT for development". Paper read at ICTD 2007 – proceedings of the 2nd IEEE/ACM international conference on information and communication technologies and development, Bangelore, 15–16 Dec 2007
Robeyns I (2005) The capability approach – a theoretical survey. J Hum Dev 6(1):94–114
Robeyns I (2006) The capability approach in practice. J Polit Philos 14(3):351–376
Robeyns I (2011) The capability approach. In: Zalta EN (ed) Stanford encyclopedia of philosophy
Sen A (1983) Poor, relatively speaking. Oxf Econ Pap (New Ser) 35(2):153–169
Sen A (1984) Resources, values and development. Blackwell, Oxford
Sen A (1985) Commodities and capabilities. North-Holland, Amsterdam/New York
Sen A (1987) On ethics and economics. Basil Blackwell, Oxford
Sen A (1990) Justice: means versus freedoms. Philos Public Aff 19(2):111–121
Sen A (1999) Development as freedom. Anchor Books, New York
Sklar A, Madsen S (2010) Global ergonomics: design for social impact. Ergon Des Quart Hum Factors Appl 18(4):4–31
Smith ML, Seward C (2009) The relational ontology of Amartya Sen's capability approach: incorporating social and individual causes. J Hum Dev Capab 10(2):213–235
Steen M, Aarts O, Broekman C, Prins S (2011) Social networking services for older people's well-being: an example of applying the capability approach. Paper read at 2011 HDCA conference, The Hague, 6–8 Sept 2011
Toboso M (2011) Rethinking disability in Amartya Sen's approach: ICT and equality of opportunity. Ethics Info Technol 13(2):107–118
Van de Poel I (2012) Can we design for well-being? In: Brey P, Briggle A, Spence E (eds) The good life in a technological age. Routledge, London
Van de Poel I (forthcoming) Translating values into design requirements. In: Mitchfelder D, McCarty N, Goldberg DE (eds) Philosophy and engineering: reflections on practice, principles and process. Springer, Dordrecht
Van de Poel I (unpublished draft book chapter) Design and well-being. Delft University of Technology, Delft
Van den Hoven J (2012) Human capabilities and technology. In: Oosterlaken I, Van den Hoven J (eds) The capability approach, technology and design. Springer, Dordrecht
Vaughan D (2011) The importance of capabilities in the sustainability of information and communications technology programs: the case of remote Indigenous Australian communities. Ethics Info Technol 13(2):131–150
Winner L (1980) Do artifacts have politics? Daedalus 109(1):121–136
Zheng Y (2007) Exploring the value of the capability approach for E-development. Paper presented at the 9th international conference on social implications of computers in developing countries, Sao Paulo

Zheng Y (2009) Different spaces for e-development: what can we learn from the capability approach? Info Technol Dev 15(2):66–82
Zheng Y, Stahl BC (2012) Evaluating emerging ICTs: a critical capability approach of technology. In: Oosterlaken I, Van den Hoven J (eds) The capability approach, technology and design. Springer, Dordrecht
Ziegler R (2010) Innovations in doing and being: capability innovations at the intersection of schumpeterian political economy and human development. J Soc Entrepren 1(2):255–272
Ziegler R, Karanja B, Dietsche C (forthcoming) Toilet monuments: an investigation of innovation for human development. J Hum Dev Capab (on-line first 4 July 2012)

# Mediation in Design for Values

A. Spahn

## Contents



**Abstract**

Mediation is the claim that technologies have an impact on the way in which we perceive the world and on the way we act in it. Often this impact goes beyond human intentions: it can hardly be understood only in terms of "intentions of the user" or in terms of "intentions of the designer." Mediation argues that technologies have "agency" themselves and then tries to explicate the way in which technological objects and human subjects form a complex relation and constitute each other. Designers should anticipate mediation effects and can use mediation to moralize technologies. However, questions can be asked about how far the moralizing of technologies is compatible with user autonomy.

A. Spahn (✉)
School of Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: a.spahn@tue.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_9

**Keywords**
Autonomy • Mediation • Nudges • Persuasive technologies • Phenomenology • Verbeek

## Introduction

Technological artifacts have, without a doubt, a big influence on our life and on social reality. Still technologies are sometimes regarded as mere neutral tools to reach the goals of humans. A knife can be used to cut cheese or to severely hurt or kill someone. It is the user (or so it seems) who is in full control over the action he intends to do. Therefore only users can shoulder responsibility: we do not put knives into prison, but only people. However, the real picture of the relation between technology and users is more complex than the idea of neutral tools suggests. The framework of mediation suggests that technologies play a much more active role in influencing the context in which they fulfill their function. Mediation is the claim that technologies have an impact on the way in which we perceive the world and on the way we act in it. Often this impact goes beyond human intentions: it can hardly be understood only in terms of "intentions of the user" or in terms of "intentions of the designer." Mediation theorists argue that technologies have "agency" themselves and try to explicate the way in which technological objects and human subjects form a complex relation, in which they constitute each other.

This chapter introduces the idea of mediation as a philosophical attempt to account for the role that technologies play in shaping human perceptions and actions. Mediation will be placed in the context of its philosophical (postphenomenological) tradition and related to similar attempts to comprehend the active influence of technologies on human behavior. Finally, the consequences for ethical technology design will be discussed and suggestion for further research will be made.

## Mediation in Design for Values

### Philosophical Approaches to Mediation

The notion of mediation is meant to overcome a too simplistic understanding of the influence technologies have on humans. Accepting that technologies mediate our perception of the world and our action in it means going beyond the idea of *mere neutral technology*. Within philosophy of technology, there is an extended debate about the moral status of technology and whether or not technology is morally neutral (cfr. Radder 2009; Kroes and Verbeek 2014). The notion of mediation has been established in philosophy of technology to overcome the neutrality thesis and take a clear stance for the non-neutrality of technology. The idea of mediation has

originally been developed within the philosophical schools of postphenomenology and STS to be able to analyze the role that technologies play in shaping human behavior and human perception.

The main argument from this perspective is that philosophers of science and technology have long neglected the ethical role of technology and interpreted technology mainly as a functional, neutral tool to reach any given aim of human actors. In this sense technology would be morally neutral, as it is only humans who can act and decide which aims to choose and which means to employ. "Guns don't kill people, but people kill people" is a popular version of the neutral technology thesis. If one starts from the assumption of the moral neutrality of technology, ethics and ethics of design can mainly be dealt with by using traditional approaches in ethics: moral philosophers can mainly focus on human agents, ignoring to a large extent the contribution of technologies to moral and immoral human behavior.

The framework of mediation challenges this view by rejecting one key assumption that underlies the neutrality thesis of technology. This assumption is often linked to the Cartesian dualism between *res cogitans* and *res extensa*, or in the field of ethics of technology: between human agents and technological objects. According to this dualism, there is a clear distinction possible between "humans" (subjects) and "technological artifacts" (objects) in a Cartesian sense: only humans have agency and intentionality, whereas objects are passive and lack agency and intentionality. Furthermore, according to the dualistic viewpoint, both humans and technology can be defined independent from each other and both have separate essential features. Scholars in the tradition of postphenomenology and STS have strongly criticized this Cartesian dualism and suggested "mediation" in the field of philosophy of technology as a more appropriate way to analyze the relation between technology and human behavior. In this section the most influential philosophical contributions in this field will be briefly sketched, before the next section links the debate about "mediation" to related theoretical frameworks to account for the non-neutrality of technology and its impact on human behavior, actions, and attitudes.

### STS

Within STS, various attempts have been made to account for the active role of technology in human decision making. Winner has famously argued that artifacts can have politics and are thus not mere neutral tools. According to him, artifacts can have politics in two different ways. On the one hand, they can "settle an issue" of political relevance. Winner argues that, e.g., Robert Moses built the bridges at the Long Island intersection in New York on purpose very low, such that busses could not pass them, thus making it difficult for poor black people to access the recreational area (Winner 1980).[1] On the other hand, artifacts might be highly compatible

[1]This particular example of Robert Moses has been challenged; see (Woolgar and Cooper 1999; Joerges 1999).

or even require a certain division of power structures to operate properly: nuclear energy requires top-down hierarchical organization, whereas solar energy panels lend themselves to more democratic bottom-up decentralized structures (Winner 1980). Therefore, technologies are not simply neutral means to realize a given end, but carry with them a political charge.

### Actor-Network Theory

To overcome the vision of technology as neutral passive tools, Actor-network theory (ANT) goes a step further and explicitly introduces the idea of agency of non-human actors (Latour 1979; Law 1999; Latour 2005). ANT suggests the notion of "actant" to cover both human and non-human actors. This is done to illustrate that agency is often distributed over various elements of a network, including the "agency" of non-humans. To avoid presupposed ideas about the items that constitute a network, all elements should be described in the same terms. ANT therefore defends the general principle of symmetry in the network in order to overcome the Cartesian dualism between humans and non-humans. As a constructivist theory, ANT tries thus to avoid all essentialism and therefore regards also non-human elements of a network as "acting." The key idea is that humans do not act alone as independent entities, but in a broader context, in which other elements of the network play a key role in determining the outcome of any action. A bulky hotel key nob can be regarded as delegating the hotel owner's wish that the guest should not forget to return the key, into the design of the artifact. In this way, the action of returning the hotel key is the result of both the actions of the hotel guest *and* the contribution of the bulky nob that serves as a physical "reminder" due to its effect of being inconveniently large.

Madeleine Akrich has introduced the notion of a "script" to account for the impact that technologies have on human agency (Akrich 1992; Akrich and Latour 1992). Like a "script" in a theater play technologies pre-scribe to some extent the actions that should be done with an artifact in question. The speed bump has the script "approach me slowly" inbuilt, the bulky hotel key incorporates the script "I will annoy you with my heavy and unhandy nob, so please return me at the reception." Akrich analyses in how far scripts are in-scribed into the technologies by designers. In the design context, a script is an anticipation of an envisaged user of any given artifact and contains assumptions about the context of usage and the series of actions that the user is likely to perform. But also assumptions about, e.g., knowledge of the user, level of familiarity with this technology, division of work between technology and user, etc. can be "in-scribed" into technology. According to Latour designers "delegate" responsibilities to the artifact: a speed bump is supposed to make sure that car drivers are slowing down (Latour 1992).

But next to the inscription of a script by the designer, artifacts can also be (re-) used by users in very creative ways that have not been foreseen by the designer or implementer of a given technology (description). In the context of her analysis of the script, Akrich also emphasizes the influence scripts have on power division and power relation, especially in developing countries (Akrich 1992).

### Postphenomenology

The most comprehensive analysis of mediation can be found in the works of Verbeek (cfr. e.g., Verbeek 2000, 2005, 2006a) who – following the work of Martin Heidegger, Don Ihde, Bruno Latour and others – develops a systematic approach of *What things do* (2005). Whereas much previous work has focused on a theoretical understanding of the mediating role of technology, Verbeek explicitly analyses the normative dimension that has according to him been neglected in prior work (Verbeek 2006a). He continues the tradition of phenomenology, which he broadly defines as the "philosophical analysis of human world relationships" (Verbeek 2006b). In line with the above-mentioned criticism of Cartesian dualism he states:

> Humans and reality are always interrelated, phenomenology holds. Human beings cannot but be directed at the world around them; they are always experiencing it and it is the only place where they can realize their existence. Conversely, their world can only be what it is when humans deal with it and interpret it. In their interrelation, both the subjectivity of humans and the objectivity of their world take shape. What human beings are and what their world is, is co-determined by the relations and interactions they have with each other. (Verbeek 2006a)

The sharp subject-object distinction, that underlies the technologies-as-neutral-tool-thesis, is thus being rejected by the phenomenological tradition as interpreted by Verbeek. According to Verbeek mediation urges us to "rethink both the status of the object and the subject in ethical theory" (Verbeek 2008a, p. 12). However, both Ihde and Verbeek at the same time want to overcome the uneasiness and critical stance against modern technology that from Heidegger on played an important role in the continental tradition. Furthermore, whereas classical phenomenology aimed to grasp "the things themselves," postphenomenology no longer accepts given relations "between pre-existing subjects who perceive and act upon a world of objects" (Verbeek 2008a, p. 13). Postphenomenology rather aims at investigating the constitution of subjectivity and objectivity in the relation between humans and reality. This relation is not stable or given, but in fact mediated by technology (Ihde 1993; Ihde 2009; Verbeek 2008a). In this line, Verbeek argues that "[t]echnological artifacts mediate how human beings are present in their world, and how their world is present to them." Accordingly the way humans act in the world and the way in which the world is represented and constituted in human perceptions is mediated by technology. What does this mean in detail?

### Mediation of Perception

Don Ihde analyses the embodiment relation of technologies. Technologies can help humans perceive the world, without being perceived themselves. When looking through a pair of glasses, the world becomes visible, while the artifact (the pair of glasses) is not perceived itself. In this example, a technological artifact becomes as it were an extension of the human body.

However, technologies also represent reality in a way that requires interpretation – which Ihde identifies as the hermeneutic relation. A thermometer represents a part of reality: the temperature. But other than a direct sensory experience of cold or heat, this representation needs to be interpreted. When a

technological artifact provides a representation of reality, there is thus almost always a selection of which part of reality is represented and which part is not represented. Furthermore, the designer of the artifact has to make a choice how to represent different aspects of reality. In this way, there is what Ihde calls a structure of "amplification" and "reduction" at play, which transforms our perception of reality. The mere fact that only certain aspects of reality are represented amplifies their significance in the interaction with the technology, while at the same time reducing all possible other aspects of reality. This transformation can go so far that technologies help to shape what counts as real (Verbeek 2006a, p. 366). Verbeek analyses the example of obstetric ultrasound to point at the various elements of a technological mediation of perception in the case of pregnancy. Ultrasonic pictures shape the way the unborn child is given in human experience. It, e.g., isolates the fetus from the female body and represents it thus as "independent," rather than as "united with the mother." Furthermore, it puts the image of the fetus in the context of medical norms, thus emphasizing "choices" and redefining pregnancy as a medical phenomenon. In this way, Verbeek claims, this mediation of perception creates a new ontological status of the fetus that is of moral significance, as it influences human decisions (Verbeek 2008a).

Another striking example of technological mediation of perception, which is ethically highly relevant, is the representation of the battlefield or terrorist subjects in the controller display of remote controlled military drones. There is an ongoing debate whether the representation of remote surveillance contributes to a dehumanization of warfare, due to an alleged video-game-like experience, or whether on the contrary it leads to more compassion, as these observations often take very long and the suspect is seen doing daily activities like playing with his children and interacting with family and friends (Singer 2009; Wall and Monahan 2011; Royakkers and Est 2010; Gregory 2011; Sharkey 2010).

These two examples make clear that the mediation of perception plays an important role in human decision making, as the way the reality is (re)-presented influences human moral decision making. This also raises issues for the design of technological artifacts, as the way technologies represent reality is often a design choice, as, e.g., in the case of medical image systems that support doctors in making decisions about health issues (Kraemer et al. 2011). Designers should be aware of the mediating role with regard to human perception of reality. They can use the insights from mediation of perception actively in designing technological artifacts. In the example of the remote control military drone, they need to reflect on whether the interface can and should be designed such that it reduces the stress-level of the operator, or whether on the contrary the design of the human-technology-interface should avoid dehumanizing effects of remote-warfare. The application of insights from the mediation of perception in other fields of technologies might be less controversial: designers can use mediation of perception to highlight important moral aspects of, e.g., consumer choices, by creating smart phone apps that give visual feedback on ecological footprints of products, their nutrient values and other morally significant features.

### Mediation of Action

Next to the mediation of perception, technology also mediates the actions of humans. This insight builds up on the ideas of "scripts" and "agency" of artifacts, which was discussed above (see section "Actor-Network Theory"). The actions of humans are determined not only by their intentions but also by their environment. A speed bump again is a classical example in which the human action (of driving) is mediated such that it becomes an action of slow driving, thereby increasing safety. The action of speeding is made (almost) impossible. In a similar way technologies invite certain actions, while at the same time making other actions more difficult. A paper cup suggests that it should be thrown away after usage, whereas a porcelain cup is designed to be used for a longer time. Design choices that shape these technologies thus affect human actions.

It is important to see, however, that only a sub-part of the mediation of action is actually intended by the designer. In many cases, the mediation of action is in fact an unintended consequence of technology design. Mobile telephones and emails have, e.g., changed the culture of interaction and communication; microwave ovens might have led to promoting the action of regularly eating instant meals individually; revolving doors were designed to keep out the cold, but have the unintended side effect of making buildings inaccessible for people with a wheelchair or a walking stick. An important insight for design is thus that technologies are "multi-stable" (Ihde 2008) and can have various unforeseen usages, including at times the complete opposite of the originally intended usage (Tenner 1997).

One example for a mediation effect that runs counter to the original intention of the design is the "rebound effect" that is often discussed in the context of design for sustainability (Herring and Sorrell 2009): energy saving light bulbs, e.g., have been designed to save energy, but it is often claimed that, due to their low energy consumption, they led to the effect that people add lighting to prior unlighted areas of their homes and gardens, leading in fact to an increase in energy usage.

This makes the task of the designer more complex, as she needs to be aware of not only the anticipated usage of the technology but also possible unintended consequences. She needs to avoid falling into the "designer's fallacy" (Ihde 2008) that the intentions of the designer are enough to anticipate the usage of a new technology and that these intentions alone should guide the design process. In a similar vein, Verbeek has urged designers to use their creativity to perform a mediation analysis to account for the otherwise unforeseen mediation effects of new technologies (Verbeek 2006a).

## Related Frameworks for Mediation: Persuasive technologies and nudges

The insight that technologies have a fundamental impact on human attitudes and behavior has also been discussed recently in various disciplines from psychology,

human-technology-interaction, design-studies, sociology, economy and philosophy of technology, without necessarily using the mediation-framework to account for it. The debates on "persuasive technologies" and on "nudging" are two examples that can easily be related to the phenomenon of mediation in Design for Values, despite the difference in terminology (Verbeek 2006a, 2009b). Let us thus look at the notion of "nudging" (with the help of technology) and "persuasive technologies" in turn.

### Nudges

Thaler and Sunstein have introduced the term "nudge" into the debate about the possibilities to change human behavior via design of the environment in which choices take place (Thaler and Sunstein 2008). They start by criticizing the idea of the *homo economicus*. Real humans should not be viewed as making rational, well-considered choices, but in fact they often lack the capacity to make choices that would be in their best interest. Following the psychological tradition of dual process theory, they distinguish two processes of decision-making. The rational slow process is good for analyzing a problem in depths and to carefully weigh arguments. Often we do in fact, however, rely on a quick and intuitive mode of making choices, which is guided by psychological heuristics and biases. Thaler and Sunstein argue that many of these biases lead us often away from choices that we would benefit from, resulting thus in many sub-optimal results.

The claim that real humans are often poor decision makers leaves in principle two strategies open: one can try to improve the abilities of humans, or change the environment in which they make choices. The first strategy would try to educate people such that they are better capable of making "good decision" by, e.g., training their decision making skills. The second strategy decides to accept that humans often tend to make bad choices, and therefore aims to adapt the environment in which humans make choices such that it "nudges" them to make better choices. Thaler and Sunstein advocate the latter option. We might, e.g., know in general that it is good to eat healthy, but this abstract knowledge alone is often not enough to motivate us in concrete situations to make healthy eating choices. This is where nudges come in: designers can structure choices such that humans decide in their best interest after all. It might turn out, e.g., that humans eat healthier if the salad and fruits are placed more prominently in a canteen, e.g., at the beginning of the row, rather than at the end of the counter.

Thaler and Sunstein thus advocate what they call choice-architecture: designers should create environments (including technologies) such that they help humans to make better choices as judged by themselves. These "nudges" should still let people free to decide (you can still ignore the salad and go for the unhealthy chocolate), but they should make the "better" choice more prominent. One can see that the idea of "nudging" would have many consequences for technology design. At the same time, it raises worries of paternalism (see section "Moralizing Technology and Mediation in Design for Values: Research Questions from the Perspective of Ethics of Technology").

## Persuasive Technology

The term "nudge" refers thus to intentional attempts of structuring human choices such that they lead to better outcomes (as judged by the individuals). The term nudge is thus very broad; a nudge can be a tax incentive, the setting of a default option on a paper-form, the choice for opt-out or opt-in strategies, the pre-structuring of complex choices, etc. One way of nudging people is to use "persuasive" strategies and embed them into technologies. With the emergence of ubiquitous ICT, technology can actively take over the role to persuade people to change their behavior: *persuasive technologies* are technologies that are intentionally designed to change human behavior and/or attitudes (Fogg 2003; IJsselsteijn 2006). Examples include blinking warning lights in cars that remind the driver to put on the seat belt, e-coaching apps to help lose weight, and RSI programs to prevent back injury.

From the perspective of mediation, one can argue that persuasive technologies exploit the mediating role of technologies, although mediation is a broader term: mediation also captures unintended influences on actions and perceptions of the user, whereas in the case of persuasive technologies the behavior and/or attitude change is an intended effect of the designer or implementer of the persuasive technologies in question. Furthermore, "persuasion" suggests that the change in behavior or attitude is voluntary (Smids 2012) and that persuasive technology goes beyond just providing information or arguments for behavior change. Persuasion can thus be placed on a continuum between mere informing or "convincing through arguments" and "manipulation" or "coercion" (Spahn 2012).

One can therefore argue that persuasive technologies are a subclass of mediation, in which (i) the behavior and or attitude change of the user is *intended* by the designer, (ii) the means of behavior change is *persuasion*, which implies establishing a (iii) minimal type of *communication or feedback*, which is (iv) often both *evaluative* and (v) *motivating* and finally allows for (vi) a *voluntary* change of behavior. Let us explain these elements of persuasion with one example. Take, e.g., an eco-feedback mechanism in a hybrid car that gives the driver a feedback on energy consumption while he is driving by changing the background color of the speedometer. A blue color signals a fuel-efficient driving style, while red indicates sub-optimal performance. Next to this colored background little symbolic flowers grow on the display, if the driver keeps on driving in a sustainable manner for a longer period. The intention of this design is to influence the behavior (driving) and perception (make fuel-consumption visible) of the driver. This is done by communicating evaluative feedback: red signals "bad" behavior. The little symbolic flowers should motivate the driver to keep on driving sustainable. But still the user is in principle free to ignore the feedback and drive recklessly if he chooses so. In designing eco-feedback systems, designers can use the mediating effect of artifacts thus actively by trying to evoke a more sustainable behavior. In a similar vein, designers can try to encourage users to adhere to other values such as a healthy diet, by creating persuasive technologies that support people in their eating choices.

In fact, many persuasive technologies are often clear examples of using mediation in the Design for Values. Recent literature has covered the application of many new fields of persuasive technologies in various domains (e.g., mobility, health care context, individual coaching, and advertisement) and for different moral aims (e.g., health, sustainability, well-being, and safety). At the same time persuasive technologies raise similar ethical issues as mediation and nudging. It can be argued that both – nudges and persuasive technologies – form subclasses of mediation, as they can be regarded as intentional attempts of the designer to exploit the mediating effects of technology for a moral aim.

## Open Questions and Future Work

The phenomenon of mediation (including nudges and persuasion) raises many challenges, both for engineering design and for philosophical and scientific analysis of the impact of technology. Roughly one can distinguish two research areas that future research needs to address. One research area concerns the foundational philosophical work on the framework of mediation. This area concerns questions about how to develop the mediation framework further and make it more fruitful for the analysis of concrete technologies and apply it systematically to issues in the design of technologies. Also the more fundamental philosophical debate on terminology and critical engagement with the framework belong to the first research area. It can be summarized as the ongoing quest to further develop, sharpen, and extend the framework itself. It concerns thus mainly questions on the level of (theoretical) philosophy of technology.

Next to this, mediation raises many questions on the level of praxis or application of mediation in technology design. If technologies have mediating effects, how should designers deal with this insight? Who is responsible for the mediation effects, the user or the designer of a technology, or both, or no-one? Should we use mediation to moralize technologies? These questions concern thus mainly the research field of applied ethics of technology.

Let us look at both research fields in turn. Since the focus of this article is on the impact that mediation has on Design for Values, the debate of the first strand of questions will be dealt with very briefly, before turning to the questions about design methodology and moral issues surrounding mediation in Design for Values.

### Mediation and Persuasion: Methodological and Meta-Theoretical Research Questions

The mediation framework has been used to shed light on the design of concrete technologies in different domains and to advance a philosophical understanding of issues of ethical technology design (cfr. Verbeek 2008b). Swierstra and Waelbers have suggested a matrix approach for the technological mediation of morality to help designers anticipate unforeseen and unintended consequences of technologies

(Swierstra and Waelbers 2012). Recently, Dorrestijn has presented an in depth analysis of technical mediation and subjectivation (Dorrestijn 2012a; Dorrestijn and Verbeek 2013; Dorrestijn 2012b). His approach draws on Foucault's analysis to account for the theoretical and ethical issues linked to the mediating role of technologies. In a similar line, scholars have investigated the potential of persuasive technologies to change human behavior (Fogg 2003; IJsselsteijn 2006; Kaptein and Eckles 2010; Verbeek 2009b; Spahn 2013).

Despite these fruitful applications, the idea of mediation has on the other hand also received some criticism within the community of philosophers of technology, mainly due to the ascription of agency and intentionality to non-human entities. Already Winner has insisted against ANT that intentionality significantly distinguishes humans from other "things," and that this difference should be taken into account when one intends to analyze the influence of artifacts on humans (Winner 1993). Feenberg and Kaplan both argue that the condition of technology should be taken into consideration more carefully in the mediation framework, be it from a transcendentalist (Feenberg 2009) or a non-transcendentalist perspective (Kaplan 2009). In a similar vein, Waelbers critically discusses on the one hand the differences and similarities between human and technological agency and intentionality, and on the other hand the consequences that mediation has for the ethical design of technology (Waelbers 2009). Peterson and Spahn have argued to re-introduce the "neutrality thesis" of technology in a weaker form (Peterson and Spahn 2011). Illies and Meijer have developed the framework of action schemes to be able to express the phenomenon that the terminology of mediation is meant to capture without going beyond established action theory terminology, especially without ascribing agency to artifacts (Illies and Meijers 2009). Pols has attempted to capture the mediating influence of technology by linking it to Gibson's notion of affordances (Pols 2013).

Verbeek has responded to some of these and other critics, amongst others by further elaborating the notion of technological intentionality and agency (Verbeek 2006c, 2009a, 2014). These terminological debates will most likely continue to occupy philosophy of technology (Kroes and Verbeek 2014) and will give room to further develop the framework of mediation or alternatives to it. Independent of these debates, one can conclude that the fact that technologies have impact on human decision making and human behavior in various ways certainly needs to be taken into account in any framework on the philosophy and ethics of technology.

## Moralizing Technology and Mediation in Design for Values: Research Questions from the Perspective of Ethics of Technology

Within the field of ethics of technology one can also identify various questions that are in need of further research: the question of whether or not to moralize technologies; the issue of the responsibility of the designer; and practical issues of anticipating mediation effects. Most of these ethical questions concern mediation, nudging, and persuasive technologies alike. These issues are not exhaustive, but all

of them triggered an ongoing-debate and deserve attention in future work. Let us briefly look at these three questions in more detail.

The first challenge in mediation for design of values is the question of whether we should moralize technology in the first place or whether we should avoid or minimize mediation effects. As seen above, Verbeek has argued that designers indeed have a moral obligation to try to anticipate the moral consequences of the mediating role of technology to avoid negative unintended consequences. However, he goes further and argues that designers should use the mediating role of technologies actively to make artifacts more moral (Verbeek 2006a, 2011). The main argument is that all technologies have a mediating role. Technologies are thus not neutral tools but have an impact on human actions, perceptions, and (moral) decision making. We therefore do not have an option to avoid the mediating effect; we should rather accept that all technologies have mediating effects. Since this is the case, designers have some responsibilities to anticipate mediation and take it into account in the process of design. They should not try to create "neutral" technologies but rather actively use mediation to create more moral technologies.

A similar argument has been made in the debate on nudging: there is no neutral design. Every design will influence the choices that humans make. If you, e.g., plan the layout of a canteen, you have to put the healthy food in some place. Where you put it will, however, inevitably influence the choices people make in the canteen, as has been argued above. Therefore, nudges should not be seen as something to avoid, but rather as be designed in a way that they both lead to better choices and preserve freedom, by giving the user the option to overrule the nudge, if he chooses to do so (Thaler and Sunstein 2008).

This idea has, however, at the same time met some resistance. Counter-arguments can take two forms: either one denies the premise that there is no neutral design by embracing a strong or weak neutrality thesis of technology (Peterson and Spahn 2011), or one accepts that all technologies have mediating effects and come loaded with values, but argues that one could still design technologies such that they maximize not specific values such as "health," "sustainability," and "well-being," but general values such as "autonomy" and "free choice." Persuasive technologies, e.g., could be designed either as nudging the user into a desired behavior, *or* as prompting him to reflect and make a conscious choice. The mere fact that technology is not neutral and has mediating effects can thus still be seen as compatible with the idea that designers should avoid paternalism and try to maximize user autonomy and free choices (Anderson 2010; Hausman and Welch 2010; John 2011). A growing field of literature therefore tries to sketch guidelines about how to take mediation and/or persuasion into account in technology design, mainly trying to balance (individual) user autonomy on the one side and design for (social) values on the other side (e.g., Berdichevsky and Neunschwander 1999; Baker and Martinson 2001; Brey 2006b; Pettersen and Boks 2008; Verbeek 2009b; Kaptein and Eckles 2010; Spahn 2011; Smids 2012; Karppinen and Oinas-Kukkonen 2013).

A second question for applied ethics research is in how far mediation changes the distribution of responsibility between designer, user, and technology. Under the neutral technology assumption, the user is always responsible for the choices he

makes, the technology is just a neutral tool that can often be used for different purposes or cannot be used at all. The mediation framework suggests a more complex relation of the distribution of responsibility, as technologies change the perception and actions of users. Future research needs to clarify who can legally and morally be held accountable for technologically mediated behavior. Here one should distinguish between the broader phenomena of mediation that also covers unintended effects on the one side and persuasion on the other side, where the change in attitude and behavior is explicitly intended by the designers (Fogg 2003). It seems that the different ways in which technology affects both individual users and societal structures or culture could be classified beyond the broader notion of mediation. What are the exact definitions, relations, and differences of various influence types such as technological mediation, affordances, persuasive technologies, and nudges (to name a few)? Such an overarching typology is still missing, even though some initial efforts have been made to develop a coherent framework to cover the various technological influence types (e.g., Brey 2006a; Tromp et al. 2011). Such a typology in turn could help solving the responsibility question.

A final open question concerns the development of a systematic method to anticipate mediation effects in the design of technologies. One suggestion could be to link mediation analysis to other design approaches that try to overcome the isolated designer choice situation by bringing stakeholders into the design process, such as in participatory technology design or constructive technology assessment. Still the specific phenomena of mediation might require a methodological tool to help designers to take mediation effects into account. An elaborated methodological tool and a systematical reflection on the best ways to reflect on mediation in the design phase are to the best of my knowledge still missing, even though many researchers have made first attempts to be more specific about how to engage in a mediation analysis in the design phase (Swierstra and Waelbers 2012; Verbeek 2006a).

## Conclusions

Technologies are more than neutral tools; they affect the user, his perception of the world and his actions in it. Mediation offers a framework to systematically account for this impact of technology on our lives. Insights from the mediation framework can be used to benefit Design for Values in various ways. Firstly, designers must be aware of the often unintended ways in which technologies shape our life. Secondly, designers can actively use mediation to moralize technologies and help people adhere to their own ethical values or to socially shared moral convictions. Designers can use the mediation framework to go beyond the neutral tool paradigm and actively shape technologies in a morally beneficial way. This raises ethical questions about how far designers can and should go in these attempts to moralize technologies, and how the balance between autonomy and social values should be settled in technology design.

Theoretical philosophy of technology has created a rich literature on mediation and various related influences of technologies on users, such as affordances,

persuasion, and nudges. The fact that mediation theory is rooted in postphenomenology makes it a coherent and systematic approach that can account for a variety of phenomena and allows integrating them into a joint framework, while at the same time offering insights both for ethics of technology and theoretical philosophy of technology. Philosophers that are critical with regard to fundamental assumptions of postphenomenology will, however, obviously take a more critical stance toward the meditation framework. The task is up to them to develop a fruitful alternative.

## Cross-References

▶ Design Methods in Design for Values
▶ Participatory Design and Design for Values

## References

Akrich M (1992) The description of technical objects. In: Shaping technology/building society. Studies in sociotechnical change. MIT Press, Cambridge, pp 205–224

Akrich M, Latour B (1992) A summary of a convenient vocabulary for the semiotics of human and nonhuman assemblies. In: Shaping technology/building society. Studies in sociotechnical change. MIT Press, Cambridge, pp 259–264

Anderson J (2010) Review: nudge: improving decisions about health, wealth, and happiness by Richard H. Thaler and Cass R. Sunstein. Econom Philos 26(3):369–375

Baker S, Martinson DL (2001) The TARES test: five principles for ethical persuasion. J Mass Media Ethics 16(2 & 3):148–175

Berdichevsky D, Neunschwander E (1999) Persuasive technologies – toward an ethics of persuasive technology. Commun ACM 42(5):51

Brey P (2006a) The social agency of technological artifacts. In: Verbeek P-P, Adriaan S (eds) User behavior and technology development. Springer, Netherlands, pp 71–80. http://link.springer.com/chapter/10.1007/978-1-4020-5196-8_8

Brey P (2006b) Ethical aspects of behaviour-steering technology. In: Verbeek P-P, Slob A (eds) User behaviour and technology development. Springer, Berlin, pp 357–364

Dorrestijn S (2012a) The design of our own lives: technical mediation and subjectivation after Foucault. Universiteit Twente. http://purl.utwente.nl/publications/81848

Dorrestijn S (2012b) Technical mediation and subjectivation: tracing and extending Foucault's philosophy of technology. Philos Technol 25(2):221–241. doi:10.1007/s13347-011-0057-0

Dorrestijn S, Verbeek P-P (2013) Technology, wellbeing, and freedom: the legacy of Utopian design. http://purl.utwente.nl/publications/88125

Feenberg A (2009) Peter-Paul Verbeek: review of what things do. Human Stud 32(2):225–228. doi:10.1007/s10746-009-9115-3

Fogg BJ (2003) Persuasive technology: using computers to change what we think and do, The Morgan Kaufmann series in interactive technologies. Morgan Kaufmann, Amsterdam/Boston

Gregory D (2011) From a view to a kill: drones and late modern war. Theory Cult Soc 28 (7–8):188–215. doi:10.1177/0263276411423027

Hausman DM, Welch B (2010) Debate: to nudge or not to nudge. J Polit Philos 18(1):123–136. doi:10.1111/j.1467-9760.2009.00351.x

Herring H, Sorrell S (2009) Energy efficiency and sustainable consumption: the rebound effect. Palgrave Macmillan, Basingstoke [England]/New York
Ihde D (1993) Postphenomenology: essays in the postmodern context. Northwestern University Press, Evanston
Ihde D (2008) The designer fallacy and technological imagination. In: Philosophy and design. Springer, Netherlands, pp 51–59. http://link.springer.com/chapter/10.1007/978-1-4020-6591-0_4
Ihde D (2009) Postphenomenology and technoscience. SUNY press, Albany (N.Y.)
IJsselsteijn W (ed) (2006) Persuasive technology: first international conference on persuasive technology for human well-being, PERSUASIVE 2006, Eindhoven, The Netherlands, May 18–19, 2006: Proceedings. Berlin. Springer, New York
Illies C, Meijers A (2009) Artefacts without agency. The Monist 36(3):420
Joerges B (1999) Do politics have artefacts? Soc Stud Sci 29(3):411–431
John P (2011) Nudge, nudge, think, think: experimenting with ways to change civic behaviour. Bloomsbury Academic, London
Kaplan DM (2009) What things still don't do. Human Stud 32(2):229–240. doi:10.1007/s10746-009-9116-2
Kaptein M, Eckles D (2010) Means to any end: futures and ethics of persuasion profiling. In: Ploug P, Hasle H, Oinas-Kukkonen H (eds) Persuasive technology. Persuasive 2010. Springer, Berlin/Heidelberg/New York, pp 82–93
Karppinen P, Oinas-Kukkonen H (2013) Three approaches to ethical considerations in the design of behavior change support systems. In: Shlomo B, Jill F (eds) Persuasive technology, vol 7822. Springer, Berlin/Heidelberg, pp 87–98, http://link.springer.com/content/pdf/10.1007%2F978-3-642-37157-8_12
Kraemer F, van Overveld K, Peterson M (2011) Is there an ethics of algorithms? Ethics Inform Technol 13(3):251–260. doi:10.1007/s10676-010-9233-7
Kroes P, Verbeek P-P (eds) (2014) The moral status of technical artefacts, Philosophy of engineering and technology. Springer, Dordrecht
Latour B (1979) The social construction of scientific facts. Hills u.a, Beverly
Latour B (1992) Shaping Technology/Building Society: Studies in Sociotechnical Change. In: Bijker WE, Law J (eds) The sociology of a few mundane artifacts, MIT Press, USA, pp. 225–258
Latour B (2005) Reassembling the social: an introduction to actor-network-theory. Oxford University Press, Oxford/New York
Law J (1999) Actor network theory and after. Blackwell/Sociological Review, Oxford [England]/Malden
Peterson M, Spahn A (2011) Can technological artefacts be moral agents? Sci Eng Ethics 17(3):411–424
Pettersen IN, Boks C (2008) The ethics in balancing control and freedom when engineering solutions for sustainable behaviour. Int J Sustain Eng 1(4):287–297. doi:10.1080/19397030802559607
Pols AJK (2013) How artefacts influence our actions. Ethical Theor Moral Pract 16(3):575–587
Radder H (2009) Why technologies are inherently normative. In: Anthonie M (ed) Philosophy of technology and engineering sciences. Elsevier B.V, Amsterdam/Boston, pp 887–921
Royakkers L, van Est R (2010) The cubicle warrior: the marionette of digitalized warfare. Ethics Inform Technol 12(3):289–296. doi:10.1007/s10676-010-9240-8
Sharkey N (2010) Saying 'No!' to lethal autonomous targeting. J Military Ethics 9(4):369–383. doi:10.1080/15027570.2010.537903
Singer PW (2009) Wired for war: the robotics revolution and conflict in the twenty-first century. Penguin Press, New York
Smids J (2012) The voluntariness of persuasive technology. In: Magnus B, Ragnemalm EL (eds) Persuasive technology. Design for health and safety, vol 7284, Lecture notes in computer science. Springer, Berlin/Heidelberg, pp 123–32. http://link.springer.com/chapter/10.1007/978-3-642-31037-9_11

Spahn A (2011) Moralische maschinen. Proceedings XXII. Deutscher Kongress für Philosophie Doc-type: Ludwig-Maximilians-Universität München (e-pub) conference object. http://epub.ub.uni-muenchen.de/12596/
Spahn A (2012) And lead us (not) into persuasion? persuasive technology and the ethics of communication. Sci Eng Ethics 18(4):633–650
Spahn A (2013) Moralizing mobility? persuasive technologies and the ethics of mobility. Transfers 3(2):108–115. doi:10.3167/TRANS.2013.030207
Swierstra T, Waelbers K (2012) Designing a good life: a matrix for the technological mediation of morality. Sci Eng Ethics 18(1):157–172. doi:10.1007/s11948-010-9251-1
Tenner E (1997) Why things bite back: technology and the revenge of unintended consequences. Vintage Publishers, New York
Thaler R, Sunstein C (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press, New Haven
Tromp N, Hekkert P, Verbeek P-P (2011) Design for socially responsible behavior: a classification of influence based on intended user experience. Design Issues 27(3):3–19. doi:10.1162/DESI_a_00087
Verbeek P-P (2000) De daadkracht der dingen: over techniek filosofie en vormgeving. Boom, Amsterdam
Verbeek P-P (2005) What things do: philosophical reflections on technology, agency, and design. Pennsylvania State University Press, University Park
Verbeek P-P (2006a) Persuasive technology and moral responsibility: toward an ethical framework for persuasive technologies. In: Persuasive technology 2006, Eindhoven University of Technology, The Netherlands. Available from: http://www.utwente.nl/gw/wijsb/organization/verbeek/verbeek_persuasive06.pdf (accessed 29 January 2014)
Verbeek P-P (2006b) Materializing morality: design ethics and technological mediation. Sci Technol Hum Value 31(3):361–380
Verbeek P-P (2006c) Acting artifacts. In: Verbeek PP, Slob A (eds) User behavior and technology development: shaping sustainable relations between consumers and technologies, Springer, vol 53, pp 53–60
Verbeek P-P (2008a) Obstetric ultrasound and the technological mediation of morality: a postphenomenological analysis. Human Stud 31(1):11–26. doi:10.1007/s10746-007-9079-0
Verbeek P-P (2008b) Cyborg intentionality: rethinking the phenomenology of human–technology relations. Phenom Cogn Sci 7(3):387–395. doi:10.1007/s11097-008-9099-x
Verbeek P-P (2009a) Let's make things better: a reply to my readers. Human Stud 32(2):251–261. doi:10.1007/s10746-009-9118-0
Verbeek P-P (2009b) Ambient intelligence and persuasive technology: the blurring boundaries between human and technology. NanoEthics 3(3):231–242. doi:10.1007/s11569-009-0077-8
Verbeek P-P (2011) Moralizing technology: understanding and designing the morality of things. University of Chicago Press, Chicago
Verbeek P-P (2014) Some misunderstandings about the moral significance of technology. In: Kroes P, Verbeek P-P (eds) The moral status of technical artefacts, vol 17, Philosophy of engineering and technology. Springer, Netherlands, pp 75–88, http://link.springer.com/chapter/10.1007/978-94-007-7914-3_5
Waelbers K (2009) From assigning to designing technological agency. Human Stud 32(2):241–250. doi:10.1007/s10746-009-9117-1
Wall T, Monahan T (2011) Surveillance and violence from afar: the politics of drones and liminal security-scapes. Theor Criminol 15(3):239–254. doi:10.1177/1362480610396650
Winner L (1980) Do artifacts have politics? Daedalus 109:121–123
Winner L (1993) Upon opening the black box and finding it empty: social constructivism and the philosophy of technology. Sci Technol Hum Val 18(3):362–378
Woolgar S, Cooper G (1999) Do artefacts have ambivalence? Moses' bridges, Winner's bridges and other urban legends in S&TS. Soc Stud Sci 29(3):433–449

# Modeling for Design for Values

Sjoerd D. Zwart

## Contents



**Abstract**

This chapter addresses societal implications of models and modeling in engineering design. The more standard question about well-known technical and epistemic modeling values, such as safety and validity, will be left to the standard literature. The sections "Introduction" and "Values in Modeling: Framing and Standard Views" discuss relevant societal norms and values and the ways in which they are model related. Additionally, standard points of view are discussed about the value-ladenness of models. The section "Value-Related

S.D. Zwart (✉)
TU Eindhoven, Eindhoven, The Netherlands
e-mail: s.d.zwart@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_1

Issues Emerging in Model Building and Use" shows various ways in which engineering models may turn out to have unforeseen societal consequences. An important way to avoid such consequences and deliberately model for values in a positive sense is to take models as special kinds of artifacts. This perspective enables modelers to apply designer methods and techniques and view a modeling problem as in need of an explicit list of design specifications. Doing so, modelers may apply forms of stakeholder analysis and participatory design. Additionally, they may apply well-known, hierarchical means-end techniques to explicate and operationalize the relevant values; doing so, they support discussions about them within and outside the design team. Finally, the model-as-artifact perspective stimulates modelers to produce technical documentation and user guides, which will decrease the negative effects of improper use. The chapter ends with a checklist of issues, which the documentation should cover if a modeling for values is taken seriously.

**Keywords**

Model • Value-ladenness • Instrumental and derivative values • Engineering, modeling, and societal and environmental values • Accountability • Affordance • Model as artifact • Modeling practices • Participatory design, value identification, and implementation • Value hierarchy • Model documentation

## Introduction

In (2002), Jean-Pierre Brans encouraged all operations research (OR) professionals to take *The Oath of Prometheus*, which is his version of the Oath of Hippocrates, well known in the medical tradition. After having done so, the OR modeler as decision-maker should not only try to achieve her or his own private objectives but should also be committed to "the social, economic and ecological dimensions of the problems." These objectives should be met "within the limits of sustainable development." Moreover, the modeler should refuse to prove "information or tools, which in [her/is] opinion could endanger the social welfare of mankind and the ecological future of Earth." This imperative of Brans is closely related to avoiding Robert Merton's third possible cause of "unanticipated consequences of purposive social action," viz., the "imperious immediacy of interest," which will be discussed in the section "How to Identify and Address Value-Related Modeling Problems." On the engineering side, the *National Society of Professional Engineers* (NSPE) expects its practitioners to exhibit the highest standards of honesty and integrity and act under the highest principles of ethical conduct. As engineering has a direct and vital impact on the quality of life for all people, it "must be dedicated to the protection of the public health,

safety, and welfare." According to the *American Society of Mechanical Engineers*, integrity and ethical conduct are core engineering values, just as are the diversity and respect, the dignity, and the culture of all people. According to engineering codes, practitioners should nurture and treasure the environment and our natural and man-made resources.

To fulfill the expectations of engineering societies, this chapter does not follow Brans. It will not formulate an Oath of Epimetheus or Hephaistos. Instead, it concentrates on forging a common ground for unforeseen value-related issues regarding model construction and its use on the one hand and values in engineering design on the other. To fulfill the NSPE requirements, for instance, design engineers should have an idea of where to look for value-related issues, and they should know how to identify and manage them. The purpose of this chapter therefore is to help these modelers to get to grips with these underexposed questions about value-ladenness of engineering models. Modeling for values in engineering design as just sketched is a large subject, and it will be delimited as follows. First, the chapter will not sketch an overview of how to achieve standard modeling values, such as verification, validation, robustness, etc. Regarding these well-known subjects, it will refer to the standard literature. Second, it does not embark upon surveying the literature on classical engineering values such as safety, risk, reliability, costs, and security. Third, we will not go into ethical questions about whether some explicit modeling purpose is morally acceptable or not. That is a general ethical question, which is not the topic of this chapter. Here, it is assumed that the purpose of the engineering model is morally acceptable.

Instead, this chapter embarks upon the questions of how to identify and solve the more hidden societal and environmental implications of modeling in engineering design. Answering these questions serves the purpose of helping model builders and users to develop more explicit ideas about the value-ladenness of model production and use. The latter involves topics like which kinds of value-related issues possibly emerge in engineering models, where to look for them, and how to address them proactively in the modeling process. To achieve this end, in the section "Values in Modeling: Framing and Standard Views," we analyze the most relevant ideas and introduce some standard positions regarding the value-ladenness of models. Next, in the section "Value-Related Issues Emerging in Model Building and Use," we will discuss some empirical findings. They present examples of unanticipated value-related issues in the practices of engineering design and how these values emerge in model construction and use. Then, in the section "How to Identify and Address Value-Related Modeling Problems," we will discuss how to handle these values in a responsible way. The main advice in this chapter is to view models as special kinds of artifacts. Consequently, the method advocated here to design for values will be to take advantage of existing design methodologies while modeling. Here, we will consider, for instance, the four-phase design cycle to operationalize, put into effect, and document model-related values in a systematic way.

## Values in Modeling: Framing and Standard Views

### Models and Modeling in Engineering Design

Models come under all forms and sizes. Almost anyone can take almost everything to be a model of anything else. This is perhaps the reason that until today all endeavors to provide an explicit definition of a model using necessary and sufficient conditions have failed. In their "Models in Science" lemma in the *Stanford Encyclopedia*, Frigg and Hartmann (2012) do not even try to give a definition. Morgan and Morrison (1999) confess: "We have very little sense of what a model is in itself and how it is able to function in an autonomous way" (p. 8). In this chapter, models in engineering design are supposed to be (1) approximate (2) representations of the target system, which is the actual or future aspect of reality of our interest. Moreover, it is assumed that models are constructed and used for (3) an explicit goal that not always needs to be epistemic. Models may be used for constructions, for explorative purposes, for making decisions, for comparison, etc. In this chapter, "model" is taken to be a notion of family resemblance such as "game" or "science" for which the three characteristics mentioned are important ingredients.[1] This model concept does not cover all the ways in which the model notion is used. Notably, people use the "model" notion in explorative contexts in which the representation element is less explicit such as in artificial life models or agent-based models. Sometimes "model" seems even to refer to something similar to a paradigm. This chapter does not cover these uses of the word model. Although the above model characterization may seem conservative, it nevertheless emphasizes the purpose of a model, which traditional engineering definitions often seem to ignore. Take, for instance, the IEEE 610 Standard Computer Dictionary. It defines a model as "[a]n approximation, representation, or idealization of selected aspects of the structure, behavior, operation, or other characteristics of a real-world process, concept, or system" (Geraci 1991, p. 132).

Embarking on the questions regarding values in modeling in engineering design requires some explicit framing of the relation between values, models, artifacts, their authors, and users. To sketch my frame, I start with a one-level description in which an artifact is produced and used. On this level, the artifact is the object (or its description) that comes out of a successful design process. The most obvious way values come about on this level is through the goals of the artifact: is the artifact made to promote the good or to inflict harm? Values also come in, however, because an artifact is the outcome of a design *process*, and this outcome is applied in society. Questions therefore arise, for example, about the

[1]Note that according to my characterization, a mathematical model "is not merely a set of (uninterpreted) mathematical equations, theorems and definitions" (Gelfert 2009, p. 502). They include their interpretation rules that define the relation between the equations and some features of the target system. "Mathematical model" is therefore a thick concept.

designers properly considering all the relevant stakes or about the users neglecting the user plan and inflicting (un)intentionally societal harm by way of the artifact's unintended application.

To finish the conceptual frame, I propose models to be *special kinds of artifacts*, which, in engineering design, aim at being applied to the design of another artifact. This results in a two-level description. The modeler produces a model, which will be used by the engineer, the user of the model, to produce a second artifact that again will be applied in society. In such a process, values come in various ways. We should consider the intrinsic values of the model and artifact and the instrumental values related to the production and the use of the model and the artifact.[2] Moreover, the situation becomes even more complicated if we realize that the second artifact could be again a model to produce still another artifact or the model may be used for more than one artifact. I will leave these possibilities out as they can be reduced to the previous situation.

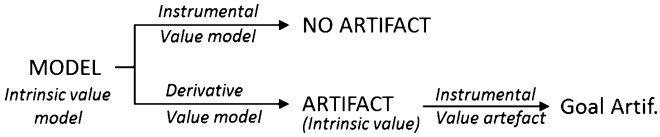
The two-level description reveals the complexity of the relation between values, modeling processes, and the products of these processes. Users apply models as a means to achieve some end, which again may be the construction of another artifact, or even a model, with again different values. The cascade of means-end relations introduces a gamut of values; in addition, all the model and artifact means-end relations are open to questions about collateral damage of the model or artifact and about more efficient ways to solve the design problem. Moreover, the actions of all the modelers and designers are amenable to normative assessments as well. In this chapter, we will observe that the traditional idea according to which professionally appropriate modeling will automatically produce morally correct models does not hold. The model-as-artifact perspective in combination with the means-end cascades undercuts this idea.

Two-level descriptions help to disentangle the ways in which models and values are related. Let us first consider the case where models are constructed as artifacts for their own sake, the first-level description. Models, then, are designed for some specific purpose and they should come with a manual. We may, therefore, at least distinguish between the *intrinsic* and the *instrumental* values of the model. The first may be historic, symbolic, esthetic, or financial values, etc., since they make the model good in themselves, and the second relate to the purpose of the model, such as decision-making, exploration, communication, simulation, etc. If on top of that a model is developed to design an artifact, the design of the model should be distinguished from that of the artifact. In such a case, we should consider at least three different ways in which values and models are related.

Like in the first-level case, in second-level descriptions, a model may have intrinsic and instrumental values, which relate to (and sometimes may even equate the intrinsic values of) the designed artifact. Third, however, and this is new in

[2]In this chapter, I will adopt Frankena's (1973) definition of intrinsic and instrumental values. The first are "things that are good in themselves or good because of their own intrinsic properties," and the last are "things that are good because they are means to what is good" (p. 54).

comparison with the first-level case, the instrumental values of the *designed artifact* often become distinctive consequences of the model with the aid of which this artifact is constructed. These may therefore be called the model's *derivative values*. Consider the paradigmatic example of the internal combustion engine pollution designed by means of a model that relates and predicts the parameters of this engine (and its polluting features). The modeling for values in engineering design explored in this chapter mainly concerns the model's *instrumental* and *derivative values*. The model's normative impact is primarily considered due to its own instrumental values and to the instrumental values of the artifact it helps to develop.




## Various Values

What are the values this chapter does focus on? Interestingly, the two codes in the Introduction put the values of honesty, integrity, and ethical conduct at the top of their lists. If we cannot count on modelers and engineers to respect these values, the discussion of modeling for values would not even get off the ground. Assuming these personal attributes of the main actors, we discern the following values among those that are generally seriously taken care of within engineering practices: *safety*, *risk*, *reliability*, *security*, *effectiveness*, *and costs* of the artifacts. I will call them the *engineering values*. Within the professional model building practices, among the values explicitly recognized are, at least, *verification*, *validation*, *robustness*, *effectiveness*, and *adequateness* of the model; I will refer to those by the term *modeling values*. As we take models to be special instances of artifacts, the first-level modeling values should directly serve the engineering ones, which is indeed the case. Many of them are directly related to the value reliability. Within a second-level description, the derivative values of a model also (indirectly) concern the instrumental values of the artifact that is based on that model such as this artifact's safety, reliability, effectiveness, and costs.

Values less universally taken into account in the technical practices of modelers and engineers mainly cluster around three subjects: *the quality of individual life*, *of social life*, and *of the environment*. The first concerns the health and well-being of human beings (and animals), their freedom and autonomy, and their privacy. More specifically, even the user-friendliness of artifacts falls within this cluster. The second cluster of values involves the social welfare of humankind, protection of public health, equality among human beings, justice, and the diversity and dignity of the cultures of all people, and so on. Finally, the third set of values clusters around our natural (and even artificial) environment and concerns sustainability and durability, sustainable development, and the ecological future of the earth such that we should "nurture and treasure the environment and our natural

and man-made resources." Let us call these three clusters together the *societal and environmental values*.

As this chapter is concerned with the values related to models and modeling, let us consider the instrumental and derivative values of models in engineering design. First, we may consider the *technical qualities* of a model in isolation without considering the contents of its purpose. The model should be built to serve its purpose without breaking down. An important quality discussed in the literature at length is, for instance, the models' *verification*. For a model as a set of equations, the latter implies, for instance, that these equations are dimensionally homogeneous, or for computer models this quality means the model should not have bugs. Other important technical model qualities are, for example, the model's *robustness* – the model should behave smoothly under small external disturbances – or *efficacy*, which is the model's ability to produce straightforwardly the desired effect. Besides these sheer technical properties, we can take the model's goal into account. If the latter is epistemic, an important *epistemic value* is its *validation*, which in many contexts comes down to the question whether the model gives an approximately true account of reality. And to what extent the model's predictions are approximately true determines its *accuracy*. Traditional modelers would probably also count the *objectivity* of a model as one of its epistemic values. The technical and epistemic values of models are first-level properties of the model itself and have been extensively investigated and described in the standard literature on models and modeling.[3] The purpose of this chapter is to help model builders and users with issues of value-ladenness of model production and use. Consequently, regarding questions about first-level technical and epistemic values, I can with good conscience refer the reader to the standard literature.

A problem less frequently addressed in this literature and therefore part of this chapter is, for instance, how different values should be weighed against each other, provided that they are commensurable at all. For instance, how should avoidance of type I errors (claiming something is true whereas in fact it is false) be balanced against avoiding errors of type II (claiming something is false whereas in fact it is true)? In science, the first is considered much more important than the latter, but this need not be the case in societal contexts (Cranor 1990). To illustrate the technical and epistemic values of a model and the moral problem of balancing them against each other, let us consider the case of the *ShotSpotter*.[4]

The ShotSpotter is a system that detects and positions gunshots by a net of microphones and is mostly used in US urban areas with a high crime rate. It is successful in drawing the attention of the police to gunshots. Trials suggest that people hardly report gunshots to the police, while the ShotSpotter immediately

[3]See, e.g., Zeigler et al. (2000); Sargent (2005); Barlas (1996); Rykiel (1996), etc.

[4]The example is from Shelley (2011) who discusses several examples of technological design with conflicting interests.

reports the time and place of a putative shot. Central to the system is a model that aims at distinguishing gunshots from other noises. If a sound largely fits the representative characteristics of a gunshot, the sound is reported as a gunshot. The model is well *verified* if it works well at every occasion it is drawn upon and never gets stuck in the process; it is effective if it does not take too much time to produce these reactions. If the model discriminates well between the sounds of shots of firearms and other but similar sounds, it is well validated, which implies that the model avoids type I and type II errors as much as possible. Statistically, however, avoiding errors of the first type implies an increase of the errors of the second type and vice versa. So, wanting to detect every gunshot implies many false positives, and decreasing false positives causes the police to miss more real gunshots. The question of the appropriate sensitivity of the model has therefore important societal implications, and its answer is not to be found in the technical literature (Cranor 1990).

Since models are artifacts, in principle all main engineering values mentioned might become important instrumental values for models as well. In the section "Value-Related Issues Emerging in Model Building and Use," we will encounter some ways in which these values might become relevant in the modeling process. The modeler, who might even be the designer of the artifact, need not always be aware of the relevant derivative values. Note that many of these engineering values mentioned before are extensively taken care of in the standards of today's engineering practices. As with modeling values, here again I will not sketch an overview of the extensive standard engineering literature. Instead, I will refer the interested reader to this literature.[5] The same even holds mutatis mutandis for technology-implicated societal and environmental values, which may not rejoice itself at an extensive treatment in the engineering literature either. Even for these values, this chapter refrains from embarking on explaining how to model for such values explicitly. I will not help the reader in finding literature on, for instance, how to model for privacy or sustainability.[6]

As we saw, models may be related to technical, epistemic, and social/environmental values, which can be instrumental and derivative. The purpose of this chapter is now twofold. First, it is to sketch various ways in which modeling projects might be, or might become, value related in ways unanticipated by the modelers; in addition, it wants to show how projects may harbor overlooked tensions between those individual values. Second, it is to make the modelers properly address these tensions within the modeling team and possibly the users and externally with the client and other stakeholders.

[5]Such as Haimes (2005)

[6]As models are special kinds of artifacts, many chapters in the present handbook discuss the engineering, societal, and environmental values mentioned in this section and more. They provide important starting points for the standard literature I have been referring to.

## Current Ideas About the Value-Ladenness of Models

Values are generally acknowledged to play a decisive role in engineering design.[7] What role values exactly play in *modeling*, however, is still controversial. Scarce attention has been paid to the question of the value-ladenness of models in engineering design. This lack of interest is remarkable as soon as one considers the massive social impact of technology and the important role values play in engineering design, which is the heart of technology. Because of the scarcity in the specific literature, we will first discuss some opinions about the role of values found in the more general modeling literature.[8]

Despite its limited size, the relevant literature displays many different opinions about the value-ladenness of models. It displays outright deniers, more cautious admitters, and militant champions of the idea. To start with the first category within the context of operations research, the idea of objective and value-free models is, for instance, expressed by Warren Walker. He maintains that the "question of 'ethics in modeling' is really a question of quality control, . . . [and] . . . as analyst . . . the modeler must make sure that the models are as objective and value-free as possible" (1994, pp. 226–227). More recently, Walker claims, ". . . if applied operations researchers (acting as rational-style model based policy analysts, and not as policy analysts playing a different role or as policy advocates) use the scientific method and apply the generally accepted best practices of their profession, they will be acting in an ethical manner" (2009, p. 051), and he argues "that the question of ethics in modeling is mainly a question of quality control" (2009, p. 1054). A similar way to go is to maintain that models themselves are value-free, whereas their alleged value-ladenness is attributed to their goal. Kleijnen (2001), for instance, claims: "a mathematical model itself has no morals (neither does it have -say-color); a model is an abstract, mathematical entity that belongs to the immaterial world. The purpose of a model, however, does certainly have ethical implications" (2001, p. 224).

Not all operations research (OR) colleagues of Walker and Kleijnen agree. Marc Le Menestrel and Luc Van Wassenhove, for instance, are cautious admitters. In (2004), they distinguish between the traditional "ethics outside OR models" just described and the more modern and radical "ethics within OR models" where the various goals of the models are mutually weighted using multiple-criteria approaches. But because "there will always remain ethical issues beyond the model," they opt for an "ethics beyond OR models" (2004, p. 480). By allowing and combining both quantitative and qualitative modeling methods, they argue that "analysts can adopt an objective approach to OR models while still being able to give subjective and ethical concerns the methodological place they deserve.

[7]See, for instance, Pahl and Beitz (1984); Pugh (1990), Jones (1992); Roozenburg and Eekels (1995); Cross (2008).

[8]Relevant literature originates in investigations into ethics in operations research and in values in computational models.

Instead of looking for a quantification of these concerns, the methodology would aim at making them explicit through a discursive approach" (p. 480). Doing so, Le Menestrel and Van Wassenhove maintain that "we should make [the need for close and on-going communication between the model builder and user] explicitly part of the [modeling] methodology" (2004, p. 480).

According to some authors, Le Menestrel and Van Wassenhove do not go far enough and disagree with them about the strength of their argument that "there will always remain ethical issues beyond the model." According to these proponents of the militant champions, we should opt for an unconditional "ethics within model" and acknowledge that models are inherently value-laden and that we should accordingly. According to Paul McNelis, for instance, "macroeconomic modeling ... must explicitly build in and analyze the variables highlighted by populist models, such as wage and income inequality ..." (1994, p. 5). Or as Ina Klaasen succinctly expresses the same issue: "Models are not value-free: moreover, they should not be" (2005, p. 181). From the perspective of science overall, Heather Douglas takes even a firmer stance and argues: "that because of inductive risk, or the risk of error, non-epistemic values are required in science wherever non-epistemic consequences of error should be considered. I use examples from dioxin studies to illustrate how non-epistemic consequences of error can and should be considered in the internal stages of science: choice of methodology, characterization of data, and interpretation of results" (2000, p. 559). More recently, she claims that "in many areas of science, particularly areas used to inform public policy decisions, science should not be value free, in the sense just described. In these areas of science, value-free science is neither an ideal nor an illusion. It is unacceptable science" (2007, p. 121).

Without going into the discussion between the deniers, admitters, and champions, let us make two conceptual observations. First, probably the meaning of the word "model" varies from one perspective to the other. Douglas and Klaasen obviously do not discuss Kleijnen's uninterpreted "mathematical entity that belongs to the immaterial world." They will consider mathematical models to be mathematical structures *with* a real-world interpretation. Moreover, these mathematical structures can mutually weigh various values, and with real-world interpretations we have values embedded in the model.[9] Second, from the deniers' perspective, the purpose of models is usually considered epistemic and models are thus similar to descriptive theories about the world much akin to Newton's model of mechanics. In this model-as-theory concept, scarce room is left for normative considerations or values, which are often considered subjective. Engineers tend to take a similar point of view. They often view models as objective representations and as such consider them part of science rather than of engineering. The advocates of value-ladenness of models conceive models however to be instruments that assist in achieving some

[9]For more on the difference between embedded and implied values in models, see Zwart et al. (2013).

(often non-epistemic) goal. This model-as-instrument conception of model is almost inconceivable without leaving considerable room for values and evaluative considerations.

From the model-as-theory perspective, one may ask why a modeler should pay attention to the ethical and value aspects of her or his creation. How could we hold Isaac Newton responsible for the V-2 rockets that came down on London and Antwerp in the Second World War? In the first place and perhaps most importantly, modelers are normatively involved in the design process because they create specific *affordances* used by the designers during this process. According to Gibson, affordances are "offerings of nature" and "possibilities or opportunities" of how to act (1986, p. 18). They are "properties of things *taken with reference to an observer*" (1986, p. 137). Gibson extrapolated the scope of affordances and applied them also to technical artifacts such as tools, utensils, and weapons and even to industrial engineering such as large machines and biochemicals. As models are artifacts and create possibilities of how to act, saying that models create affordances is clearly within Gibson's original use of the word. Affordances of artifacts, objects, or any instruments therefore can broadly be conceived as those actions or events that these artifacts, objects, or instruments offer us to do or to experience. Consequently, models afford us to get knowledge or to decide about design proposals, which perhaps even did not exist before the model was created.

Generally, we may say that creators of affordances are at least *accountable* for the consequences of these affordances. One may define accountability to be the moral obligation to account for what happened and one's role in making or preventing it from happening. In particular, a person can *be held accountable* for X, if the person has (1) the capacity to act morally right (is a moral agent) and (2) has caused X and (3) X is wrong (Van de Poel 2011, p. 39). Accountability is to be distinguished from blameworthiness. An agent can be accountable for an event but need not be blameworthy as she or he can justifiably excuse herself or himself. Typical excuses are the impossibility to be knowledgeable about the consequences of the event, the lack of freedom to act differently, and the absence of the agent's causal influence on the event (Van de Poel 2011, pp. 46–47). For instance, a manufacturer of firearms may be held accountable for the killing of innocent people. As the creator of the affordance to shoot, she or he may be asked about her of his role in the killings by the guns made in her or his company. The typical excuse of the manufacturer reads that she or he did not do the shooting and therefore is not (causally) responsible for the killing. In this sense, the creators of affordances are accountable but need not be blameworthy for the consequences of these affordances. Similarly, but often less dramatic, modelers are accountable for the consequences of the affordances, viz., the design, because they willingly brought into being the affordances of their models. Consequently, if the capacity, causation, and wrongdoing conditions are fulfilled, modelers are accountable for the properties of the final design and may even turn out to be blameworthy and should therefore pay attention to the normative aspects or their creations.

## Value-Related Issues Emerging in Model Building and Use

In the present section, we will encounter various ways in which constructing models may have moral and societal consequences.[10] These sketches serve the heuristic purpose for model builders to create awareness about where and how to look for the normative implications of the process and product of modeling. We will see that indeterminacy of the modeling question, the boundaries of the model, underdeterminacy or the complexity of the modeling problem, lack of knowledge and uncertainty, and finally embedded value assessments in the models may have unforeseen value-laden effects. We start with the situation where even the concept of the model does not exist yet and where it is even unclear what the target system should be. Then, we turn to the possible underdeterminacy of a model, and after that we will consider complexity and uncertainty as possible sources of normativity. We will end with the necessity to explicate the purpose of a model clearly and to communicate it.

### Indeterminacy of Model Building Question and Model Boundaries

When a modeling process concerns an innovation, the initial formulation of the modeling problem is often vague and indeterminate, which means that the problem lacks a definite formulation. On top of that, even its conditions and limits may be unclear (Buchanan 1992, p. 16). Indeterminate problems have a structure that lacks definition and delineation. Herbert Simon (1973) calls these kinds of problems "ill structured." At the outset of the problem solving process, an ill-structured problem has unknown means and ends. Often, modeling problems are ill structured to such an extent that they become "wicked problems" (Rittel and Webber 1973). Wicked problems have incomplete, contradictory, changing, and often even hardly recognizable requirements. Through interactions with the client and within the modeling team, these problems usually become better structured over time. Usually, however, they may leave open many acceptable solutions. This may be due to partial knowledge or understanding of the model requirements. It may also be due to the underdeterminacy of the target system when the physical laws applied do not fix the variables involved. Helping to separate the relevant from the irrelevant aspects of the modeled phenomenon has a normative side. Besides weighing epistemic values such as determining the real-world aspects that need to appear in the model, the definition of the modeling problem also fixes the scope on the societal effects taken into account. For instance, traditional thermodynamic optimization models for refrigerator coolants favored Freon-12 because it is stable, energy efficient, and nonflammable (in the volumes used). When the model includes sustainability, however, this coolant loses its attractiveness because of its propensity to deplete the ozone layer. In 1996, the USA banned its manufacture to meet the Montreal Protocol.

[10]The examples in this section come from participatory research reported more in detail in Zwart et al. (2013).

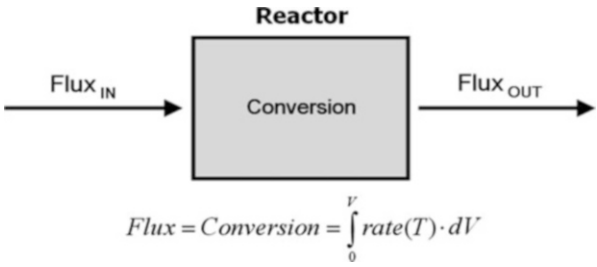



**Fig. 1** Model of chemical conversion at steady state determined by temperature T and volume V

Interestingly, questions about values still arise even at the stage in which the modeling situation is determined and the model can sufficiently accurately describe the behavior of the fixed target system. The following simplified example shows how a seemingly straightforward application of a mass balance relates to societal norms and values. In most (bio)chemical processes, the goal of the process design is the conversion of one substance into another. Usually, the design requirements fix the product mass flux, i.e., how much substance is to be produced in a given time span. Let us for simplicity's sake assume that the conversion rate of a reactor is 100 % and that conversion only depends on the reactor volume ($V$) and the reaction temperature ($T$). The steady-state mass balance of the reactor then can be modeled as is depicted in Fig. 1. Under these circumstances, clearly the modeled system is underdetermined and the design team is free to choose the reactor volume or the reaction temperature.

Despite its simplicity, the conversion case already raises interesting questions about societal consequences and thus value-ladenness. Although the modeling choice of volume and temperature is seemingly a factual affair, their trade-off has considerable derivatively value-laden implications. The larger the reactor, the larger the volume of the contained substance; the larger the possible amount of possibly spilled substance in case of leakages or other calamities, the larger the volume of substances managed during the shutdown and start-up of the plant. Moreover, extremely high or low temperatures can cause major hazards to operators and provide sustainability issues due to high energy requirements. Thus, fixing the temperature-volume region has societal implications regarding environmental issues, safety, and sustainability hazards. Does this validate the assertion that the flux model is value-laden, or should the responsibility for these value-related issues be exclusively put at the feed of the designers of the reactor – provided they are not the same persons?

Answering that some model is only a set of equations will not do. Following our characterizations of a mathematical model, the last includes rules of interpretation and therefore is more than just a set of mathematical equations. Nevertheless, the flux model does not explicitly embed a value judgment because it is silent about what combinations of volumes and temperatures are preferable. Quite the contrary, the model refrains from any direct value judgment and only describes the relation between the variables within some margins of error, and the only value involved is the model's accuracy. However, the representation of the situation as only a

physical relation between variables without safety and sustainability considerations already implies an evaluative stance. The model could have incorporated information about unsafe volumes and energy consumption. Thus, the choice of the model's constitutive parts and the absence of reasonable upper and lower limits render the model value-laden in the derivative sense. The absence of societal aspect reflects the modelers' judgment that they are not important enough to be considered. From the present considerations, we may conclude that the delineation of the modeling problem and the decision to put boundaries on the values of the model's variables or not have social and environmental consequences.

## Underdeterminacy of the Physical Situation

Even if the target system of a modeling procedure and the models' boundaries are fixed by the design context, the model may be underdetermined, and underdeterminacy may also be a source of unnoticed normativity. Consider, for instance, the example of a model describing the output of an electrodialysis procedure used to regain acids and bases from a plant's recycled waste water stream. In Zwart et al. (2013), we observed two steps in its related model development. The first version of the model merely represented the features of the electrodialysis. It described the relation between an electric current and a membrane area (at which the electrodialysis took place) for a fixed production of bases and acids. This first version of the model, however, was underdetermined since it failed to suggest which size of the membrane or which current was to be preferred. To come to a unique working point, the modelers added economic constraints to the model.

The second version of the model included therefore also economic equations, which allowed for calculating the membrane size by reducing the total cost of the process. The newly introduced considerations had a significant impact on the model's characteristics. Whereas the first version of the model was merely descriptive, after the introduction of economic constraints, it became normative. After this introduction, the model could identify the optimal design but at the detriment of other values, such as safety and sustainability. If the modelers had considered sustainability considerations to fix the optimum, their model had perhaps come to a different preferred working point. These and similar examples show that model optimizing strategies are very likely to introduce, often unrecognized, normative issues.

## Complexity, Lack of Knowledge, or Uncertainty

Besides the indeterminacy of modeling questions or the underdeterminacy of the physical description of the problem situation, we consider three additional sources of value-related issues: the complexity of the target system, lack of knowledge, and uncertainty about the behavior of the target system. Many modeling situations in engineering design are far too complex to be handled in all details at once.

**Table 1** Differences in models' complexity

| | Decreasing constants | Decreasing variables |
|---|---|---|
| 1 | $r = k_n.[E].\frac{[S]}{k_m+[S]}.\frac{[S]^2}{k_r}$ | $r = k_t.(T).[S].\frac{1}{[P]}$ |
| 2 | $r = k_i.[E].\frac{[S]}{k_m+[S]}$ | $r = k_p.[S].\frac{1}{[P]}$ |
| 3 | $r = k_s.[E].[S]$ | $r = k_f.[S]$ |

Design engineers apply different methods to cope with these situations and many of them have normative implications. We mention some of these: reducing the number of variables and constants in the model, neglecting the starting-up and shutdown phases, and carving up the problem in more manageable sub-modules.

First, reduction of the number of variables can be done by treating them as constants, and reducing the number of constants is helpful when the theoretical value of the constant is unknown and hard to establish in a reasonable time. In such situations, the value of the parameter is estimated or, sometimes, is just left out. The reduction of variables and constants will usually introduce inaccuracies. To illustrate this phenomenon, let us consider an example in which an enzyme is used as a catalyst for a biochemical conversion. To model the reaction rate $r$, various models are available; see Table 1 in which the $k's$ are different constants; $[E]$, $[S]$, and $[P]$ are concentrations of reactants; and $T$ is the temperature. Models with fewer constants and variables (the left and right columns of Table 1, respectively) become less accurate. The former ones have a smaller range of application, whereas the latter consider fewer dependencies. Notice that both the reduction of variables and constants result in a change of the values of the constants in the reduced equation.

The resulting models are value related for at least two reasons. First, the epistemic values of *accuracy* and *generality* are assumed to be of less value than the pragmatic and non-epistemic value of being able, pragmatically, to model the target system at all. Second, if, for example, temperature is left out of the equations, it cannot be taken into account anymore, when *safety* or *reliability* of the system turn out to be temperature dependent. The decision to leave temperature out, then, entails a derivative value judgment.

A second way, in which modelers avoid complexities in the modeling process, is to concentrate only on the steady state of the process, neglecting the modeling of the system's start-up and shutdown phases. After all, static situations are much easier to model than dynamic ones. Only considering steady-state modeling, however, leads to neglect of the system's dynamic behavior and substantially decreases the model's range of application. From the viewpoint of safety, this neglect is undesirable as in practice the start-up and shutdown phases of large-scale (bio)chemical process are the most dangerous (Mannan 2005). The modeling decision to focus on the steady state and to neglect the start-up and shutdown dynamics has therefore normative implications in the derivative sense.

A third way to cope with the complexity of a modeling problem is to divide it into independent parts and try to solve the less intricate problems posed by those parts. This modularity in the modeling approach, however, sometimes poses its own hazards. In the electrodialysis example mentioned during the production of acids

and bases from salts, some hydrogen and oxygen gases were produced at the electrodes. The conversion model neglected the production of hydrogen and oxygen, because of its minor impact on the efficiency and the mass balance. At a later design stage, however, when carrying out the hazard and operability (HAZOP) analysis, the modelers failed to recognize the hazards posed by the free production of hydrogen and oxygen together. It turned out therefore that although the simplifying assumptions were taken with great care and were harmless on small scale, when the design was scaled up, they posed a much larger risk.[11] The electrodialysis example nicely illustrates the dangers of scaling within a modular approach and more generally the context dependency of value assessments.

## Proper Use of the Model and Communication

In the examples of the previous section, the model builders were in close contact with the users of the model or were even identical to them. In situations where the users of the models are unfamiliar to the modelers, different kinds of problems emerge regarding the value relatedness of models. The first issue concerns the model's use. When we take models as special kinds of artifacts, we recognize they may or may be not be used according to the modelers' intentions. The first will be called proper, and the second improper use of the model. Proper use of artifacts is closely related to a second issue, viz., appropriate communication about the model's proper use. Thus, instructions about proper use and the model's normative dimensions require effective communication. While the importance of communication between modelers and other stakeholders is hardly denied in theory, it is often neglected in practice. Let us turn to two examples illustrating problems with improper use and insufficient communication: one exhibits *instrumental* values, and the other *derivative* ones.

The first example concerns geographic information systems (GIS) as a decision support system. These systems model the topography of a landscape, representing differences in height, water streams, water flow, and the type of landscape (e.g., forests or plains). Sometimes these hydrological models are used for decision-making in geo-engineering. However, as Jenkins and McCauley (2006) describe, the use of GIS procedures may raise problems because of the application of GIS, SINKS, and FILL routines. GIS programmers aim at approximating topographies in a simple way while keeping high data accuracy. To that end, they make two assumptions. First, they assume that rivers and water streams follow a fractal geometric structure, i.e., the whole river system has the same structure as its parts. This assumption increases the simplicity of the model, because it simplifies the recognition of river systems. According to the second assumption, local

[11]The 1991 Sleipner case shows that inattentive downscaling also can cause catastrophes. See Selby et al. (1997) for the details of how a concrete offshore platform collapsed due to incorrect downscaling of an FEM model.

topography is flat. Consequently, isolated sinks or mounds are assumed to be noisy data. This assumption increases data accuracy because often depressed or elevated cells do not correspond to reality. Unfortunately, the assumptions taken together have the tendency to sort out valuable wetlands, which "provide important ecological services such as flood mitigation, groundwater recharge, nutrient processing, and habitat for unique flora and fauna" (Jenkins and McCauley 2006, p. 278). Assumption one sorts out wetlands, because they seem unconnected to the branches of rivers and are not recognized as a part of the water stream system. Assumption two does the same as wetlands are depressed areas. In combination, the two assumptions may even sort out comparatively large wetlands. Consequently, even when geo-engineers use GIS models to aim at decisions with low environmental impact, they may unintendedly opt for destroying wetlands if they are unaware of the mechanisms just mentioned.

Jenkins and McCauley suggest several solutions (2006, pp. 280–281). In their view, GIS programmers "could try to educate users about the limitations of some of the algorithms." They may also let their programs always "produce output that more accurately reflects the actions undertaken to produce the model." It would "help the end users better understand that GIS products are no more than [just] models of real ecosystems and landscapes." The most effective solution according to the authors is the technical fix, viz., to let the programmers "change the assumptions in the current modeling algorithms to avoid assumptions that 'fill' isolated wetlands." They also contend that the "menu-driven, point and click interface," besides the more technical command line, increases the risk of accidents. "Organizations that provide GIS data layers and products (e.g., maps, databases) that include hydrological models described above should carefully examine the assumptions of the model, including ramifications and limits on the ethical use of models given those assumptions, and should prominently list those assumptions and ramifications in metadata associated with data and products." Regarding the question of who bears responsibility, Jenkins and McCauley claim that as the "GIS programmers are in a position of power," "[t]he locus of responsibility [. . .] reside[s] mostly with the GIS programmers." This assessment follows the ethical principle that a person with more power in a situation has increased responsibilities in that situation.

Although the GIS example just described may originate in specific concerns about the disappearing of wetlands, it nevertheless makes a clear example of the dangers produced by improper use of models. After all, the hydrological GIS models are constructed with routines for streaming and not for stagnant water. More recently, the issue of improper use of model and its ethical relevance has received more specific attention in the literature. Kijowski et al. (2013), for instance, gathered empirical information by consulting nineteen experts about their experiences with computational models. Based on this information, they discussed ways in which model builders and users may reduce improper uses of models.

The normative difficulties with the GIS example just mentioned relate to the model's own *instrumental* values, which are directly related to the goal-related consequences of the model itself. The model is used to support decision-making and not to construct new artifacts. Jenkins and McCauley describe how the long

distance between modeler and user may result in improper use with serious environmental consequences. In the next section ("Documentation About the Values"), we will see that in system engineering a long distance between designer and user (here the modeler and user) asks for extensive *database documentation*. When the lines between modeler and user are short or even nonexistent, the advice is to stick to the less extensive *self-documentation*. The combination of the two may still harbor dangers as the second example shows. It features a small distance between modelers and model users (the designers of the artifact) and a large distance between the designers and the users of the artifact based on the model – the Patriot case to be presented below. This combination results in *derivative* model values related to the safety of the artifact.

On the 25th of February 1991, a Patriot missile defense system at Dhahran, Saudi Arabia, failed to intercept a Scud rocket, due to an "inaccurate tracking calculation" in the trajectory model; tragically, 28 soldiers died and another 98 soldiers were wounded.[12] The main problem was related to the model representing the rocket trajectories. Originally, the Patriots were designed to operate against armies with a highly developed intelligence, such as the former Soviet Army. Consequently, the missile systems were assumed to have to deal with enemy detection for only short periods. In the First Gulf War, the intelligence capabilities of the enemy were less sophisticated and therefore the Patriot systems were used over several consecutive days at several sites. Over longer periods, however, the approximations of the model calculating the Scud trajectories became less accurate, and after about 20 consecutive hours of operation without reboot they became useless. The main problem, the modeler's decision to allocate insufficient bytes to represent the variable time, was not resolved before the failure happened, because the modelers "assumed . . . Patriot users were not running their systems for eight or more hours at a time" (Blair et al. 1992, p. 8), though several model updates had been made in the months before the failure. The modelers also sent out a warning that the Patriots were not to be used for "very long" periods. However, neither the users' assessments nor the instructions for use had been explicit enough to prevent the incident (Blair et al. 1992).

Probably, the Scud incident would have been prevented by more explicit and intense communication between the model builders, the Patriot designers, and its users. If the designers had consulted the users more explicitly and if the modelers had informed the designers about the uselessness of the model after 20 h of operation, the incident would probably not have happened. The Patriot case provides a good example in which improved communication and transparency would have decreased the risk of the accident significantly.

Recently, various authors make a plea for improved communication among all the stakeholders involved in the development and use of models. For instance,

[12]After the Gulf War, discussions arose about the efficacy of the Patriot defense system (cf. Siegel 2004), and the software failure was criticized for being just a scapegoat for the army to cover up the malperformance of the Patriot system. This discussion however does not subvert the example. Even if the critics are right, we may consider the Patriot software failure to be an instructive imaginary case. See for a more detailed account Diekmann and Zwart (2013).

Fleischmann and Wallace (2005) break a lance for transparency regarding the working mechanism of decision support models. They argue that "the outcome of modeling depends on both the technical attributes of the model and the relationships among the relevant actors. Increasing the transparency of the model can significantly improve these relationships." The plea for more transparency and better communication between model builders and users is taken up and discussed by Shruti and Loui (2008). There, the authors use their seventh section "Communication Between Modelers and Users" to comment and elaborate Fleischmann and Wallace's reasons to advocate transparency for model builders.

The ways in which models become value-laden and the examples discussed in this section closely relate to the subject "unintended (or unforeseen) consequences" in the social sciences. In his seminal (1936) paper, "The Unanticipated Consequences of Purposive Social Action," Robert Merton distinguishes five causes for those consequences. The first is "lack of adequate knowledge" or the factor of ignorance, which he carefully distinguishes from "circumstances which are so complex and numerous that prediction of them is quite beyond our reach" (p. 900). This factor is clearly connecting to the section "Complexity, Lack of Knowledge, or Uncertainty." Second, Merton identifies "error." He writes: "the actor fails to recognize that procedures which have been successful in certain circumstances need not be so under any and all conditions." The latter is related to not anticipating what happens with your model outside the specs, something for which we can hold the modelers accountable for the Patriot-system accident. Also the GIS example is related to the case of improper use. Jenkins and McCauley's (2006) subtitle justly reads: "unintended consequences in algorithm development and use." As a third factor, Merton mentions "imperious immediacy of interest," which he describes as "the actor's paramount concern with the foreseen immediate consequences excludes the consideration of further or other consequences of the same act." This factor is similar to the "collateral damage" we discussed in the section "Models and Modeling in Engineering Design," and we will come back to it in the section "How to Identify and Address Value-Related Modeling Problems." Besides Merton's fourth factor "basic values," which may have detrimental effects in the long term, he calls on as a fifth reason something like self-defeating prophecies. He describes them as: "Public predictions of future social developments are frequently not sustained precisely because the prediction has become a new element in the concrete situation, thus tending to change the initial course of developments" (p. 903/4). We did not encounter examples of self-defeating prophecies here, but the phenomenon of self-defeating and self-fulfilling prophecies is highly relevant for modelers building models' policy decision support.

## How to Identify and Address Value-Related Modeling Problems

This section is dedicated to the practices of modeling for values in engineering design. More specifically, we will discuss the question of how modelers may identify the most relevant instrumental and derivative values in an engineering

design modeling process. Only identifying these values however does not suffice. Modelers should have ways to find out how to manage and to realize these values. Even this is insufficient. In the end, during the evaluation process, the modeler should also be concerned about the aftercare of her or his model and the related artifact. She or he at least should take care of an adequate user plan and sufficient communication with the users about the model. How are we to structure this process of modeling for values? An important way to go is to apply more explicitly engineering *design methodologies* to modeling itself. Acknowledging models to be special kinds of artifacts with well-elaborated design specifications and specified goal will help the modeler to manage systematically the values (possibly) involved in her or his modeling assignment.

When a modeler explicitly follows the outline of some design methodology, it will support the modeling process for at least two reasons. First, it frees her or his mind from the traditional idea that the purpose of the model is always epistemic and makes her or him realize that models can, and often do, have other goals. Second, applications of proven design practices help the modeler to think outside the constraints "given by the design commissioner."

Design methodologies exist in many forms and sizes; the one we will follow is inspired by Jones (1992), Roozenburg and Eekels (1995), and Cross (2008). They take design methodologies as combinations of some distinguishable phases.[13] First, we identify a diverging phase of *analysis*, in which the design brief is analyzed, redefined, and perhaps subdivided in subproblems and in which the design specification is determined. In this phase, the goals are set and the design specifications are operationalized. The second is a transforming *synthetic* phase where the working principle is chosen and the design is built up from its parts; it results in a provisional design, models, or even prototypes. Third, with the provisional solution at their disposal, in the *simulation* phase, the design team finds out, through reasoning or experimentation, how much the prototype exhibits the required behavior. Fourth, in the *evaluation* phase, it is decided whether the prototype fulfills the requirements well enough or whether it requires optimization or even replacement by a new design proposal. If the second holds, the design cycle should be applied again until the outcome satisfies the design specifications. When the design process has finished, the design team should communicate about the design with the outside world and provide extensive technical documentation.

The application of the four-phase design cycle to model construction has two important consequences. First, a model should come with a design brief and design specification; its purpose and functionalities should be stated, such as its limits, the preconditions of working conditions, and its domain of application. Its design specification should operationalize the most important requirements including all its *instrumental values*. Second, when a model is developed for engineering

[13]See also the ABET (1988) definition of design, which states "Among the fundamental elements of the design process are the establishment of objectives and criteria, synthesis, analysis, construction, testing and evaluation," or the ISO (2006) section 5.

design, some members of the modeling team should also participate in the design cycle of the artifact for which the model was built. Doing so, the modeling team can be sure that the model design specifications cover the most important *derivative* values and the model adequately manages the value-related issues in an artifact design.

Focusing on model values, the next subsections loosely fit the phases just described above. The section "Identifying and Accounting for Instrumental and Derivative Values" instantiates the analysis phase in which the model design brief should be specified and the most important instrumental and derivative model values should be identified and operationalized. Similar to the synthetic phase, the section "Operationalization and Implementation of Values" reports on one possible way to operationalize different values in a model building context. It discusses Van de Poel's (2013) values hierarchy. Finally, parallel to the evaluation phase, the section "Documentation About the Values" will discuss the aftercare for the model once the modeling is done. It stresses the importance of documentation and communication of all the value-related issues concerning the final outcomes of the modeling process.

## Identifying and Accounting for Instrumental and Derivative Values

The application of the design-cycle perspective to the modeling process with the focus on values yields interesting observations. Let us start with the model design brief, which states the model design problem. First, this commission should explicate the *modeling problem itself*, the *owner* of this problem, and it should clearly state *the goal of the model*. Second, the commission should also *explicate the context of application*. If we consider the *instrumental* values, the commission should elaborate on the values directly involved in the model purpose and on the values more indirectly involved in the modeling problem. Because making a model is an action, the modeling team should list all the possible answers to the modeling problem and should proactively think about the possible "collateral damage" of the various actions and models involved. Moreover, the commission should at least identify and discuss the possible tensions and incompatibilities between the values involved. To come to the *derivative* values of the model, the same questions have to be asked about the artifact to be designed.

When we turn to the model design specifications, the application of the design cycle again provides important value-related insights regarding the model's instrumental and derivative values. Overall, the design specification elaborates the objectives of the object or process to be designed. It is the list of criteria fixing the extent to which the designed object fulfills its purpose. The design specification is therefore the *main* instrument for assessing possible design proposals; this is its main function. Applied to modeling for values, the model design specification is the most appropriate place to state and elaborate to what extent the model satisfies its instrumental and derivative values. Regarding the model's instrumental values, the list should specify its technical values and, if its goal is knowledge,

its epistemic ones. Even if the model does not serve the purpose of constructing an artifact, this does not suffice. Stand-alone models often have societal and environment consequences; the GIS case with the SINKS and FILL routines provides a telling example. And as the Oath of Prometheus, mentioned in the Introduction, explicitly states, the modeler or modeling team should also think proactively about all what we called the societal and environmental values possibly involved in their models.

When the aim of a model is the construction of a technical artifact, the modeling team should at least consider artifact design brief and design specification to come to the relevant *derivative* values. Moreover, it should participate in the various iteration of the artifact design cycle to keep track of changes in the artifact design specification and values involved. In the electrodialysis example, this participation and the attempt to put all values concerned in the model design specification would have revealed that the introduction of costs could possibly be at the detriment of societal and environmental values, and in the ShotSpotter example it would have explicated more clearly the weighing of type I against type II errors. In the next subsection, we will return to the question of how to organize the values within a model design specification.

As a design specification needs to be as complete as possible, application of the design cycle to modeling for values implies the quest for a list of instrumental and derived modeling values that is as complete as possible. Moreover, when a supposedly complete set of those values has been gathered, the question arises of how these values should be weighed against one another. Here again I would like to draw upon the inheritance of design practitioners. To come to a complete-as-possible list of values and decisions about how to weigh them, modelers should follow designers who, in various democratization movements, consult the *most relevant stakeholders* to establish the design specifications and their relative weights.

One of these movements originated in Scandinavia of the 1970s and was driven by the urge to let the users cooperate with the designers within the design process. For that reason, it was called *cooperative design*. When exported to the USA for political reasons, the name was changed to *participatory design* (Schuler and Namioka 1993). Participatory designers advocate an active involvement of all stakeholders (such as designers, clients, regulators, users, fundraisers, and engineering firms).[14] Science and technology studies have been another breeding ground for the appeal for increased democratization in technological design. After allegedly having "falsified" the traditional perspective that technology has a determining influence on society, sociologists of technology claimed to have shown that the successful functioning of artifacts is a *social construction*. Along this constructivist line of thinking, they advocated more participation of citizens in

[14]For recent developments in participatory design, see the special issue of *Design Issues* on the subject, volume 28, Number 3, Summer 2012, or the proceedings of the biennial *Participatory Design Conference* (PDC), which has had its 12th meeting in 2012.

technological developments and design (see, e.g., Bijker 1995; Feenberg 2006). The developments in the participatory-design and democratization process came together in a special issue of *Design Issues* in the summer of 2004 reporting on the symposium "An STS Focus on Design." In this issue, Dean Nieusma writes: "participatory decision making is (1) fairer and (2) more intelligent than nonparticipatory processes." To show participatory design is fairer, Nieusma cites Schuler and Namioka (1993, p. xii) who say: "people who are affected by a decision or event should have an opportunity to influence it."

To elaborate the last statement, Diekmann and Zwart (2013) interpret the democratization movements in modeling and design to be valuable steps into the direction of *modeling for justice*, i.e., reaching an overlapping design consensus possibly with all the stakeholders involved. This consensus provides a foundation for value decisions that is morally more justified than just letting the modelers and designer balance the values they identified or elaborate cost-benefit analyses. In the same vein, Fleischmann and Wallace discuss in (2005) the stakes of the "various actors involved in decision support: the modelers, the clients, the users, and those affected by the model" (see also Fleischmann and Wallace 2009).

A difficult but unavoidable question here reads: But who are the relevant stakeholders?[15] Overall, stakeholders are those who have an interest in the design, such as customers, consumer organisms, producing companies, suppliers, transporters, traders, government, etc. However, who is to decide, who has a genuine interest, and whether an alleged interest is important enough to establish a stakeholder? Not every self-proclaimed interest suffices to qualify as a relevant stakeholder. After all, is someone who wants all airplanes to be yellow a stakeholder in airplane design? What shall we decide about future stakeholders? Is a healthy person who will be seriously sick in 10 years' time a stakeholder in the design of medicine? Are future state citizens stakeholders in the decision about a new nuclear energy plant? Questions arise about voluntariness as well. Parties that are involuntarily affected by the design may rightfully be called stakeholders. Consequently, pedestrians are stakeholders in car design and perhaps even more than the future car owners. After all, the customers can refrain from buying the car, whereas pedestrians are potential victims in accidents and cannot choose whether they want to be hit by the car.[16]

Besides consulting stakeholders, designers also carry out life cycle analyses to complete the design specification, and they consult standard checklists, such as those of Hubka and Eder (1988, p. 116), Pahl and Beitz (1984, p. 54), and Pugh (1990, pp. 48–64). These lists may also be useful for finding unidentified model values. The identification of these values is important but only a necessary

[15]Woodhouse and Patton (2004, p. 7) ask a similar question within the STS context of design: "Who shall participate in making decisions about new design initiatives (and in revising existing activities)?"

[16]Finding out how to identify the relevant stakeholders and their views, modelers could also explore the way system and software engineers carry out *requirement analysis*, which covers among other things stakeholder identification and joint requirement development sessions.

condition for a satisfactory value design specification. In the next section, we will come to the question of how the various values could be organized in a design specification to come to an operationalized set of values.

## Operationalization and Implementation of Values

Although the completeness of the design specification regarding values is an important necessary condition for modeling for values, it is not sufficient. A large set of partially interdependent values, norms, and design specifications without any structure would be very impractical and too difficult to manage and adjust during the modeling and design process. In contrast and addition to the completeness of the model-related set of values, this set should avoid redundancy and promote independencies among its values. Besides the tension between completeness and nonredundancy, the third quality constraint for the values in the model's design specification is the clearness of their meaning and the way they are operationalized. As values are often abstract concepts, their meaning in a specific context should be explicated, such that the extent to which the model or design fulfills the value design criteria can be assessed intersubjectively. In other words, the value criteria in the design specification should be testable. Finally, the modelers should take care that the proposed way to operationalize the abstract values are valid, that is, whether they still carry largely the same meaning of the abstract values they started with at the outset. To serve the purpose of nonredundancy, appropriately operationalized values, and validity, we will consider Van de Poel's (2013) method of values hierarchy.

To come to a set of valid, intersubjectively operationalized and testable, complete but nonredundant design requirements, the design literature often uses the instrument of a *hierarchical tree of design objectives* (e.g., Cross 2008, pp. 65–71; Roozenburg and Eekels 1995, pp. 141–143). At the top of those trees, the most general and abstract objectives of the artifact are situated, and the lower nodes refer to subgoals that should be reached to serve the final goal at the top. According to Cross (2008), for instance, an intermediate means serving the goal of a safe new transport system is: "a low number of deaths." This last intermediate objective is again served by the means of a "high speed of medical response to accidents" (p. 69). The objective trees or means-end hierarchies normally contain various branches and many nodes where the lower nodes are the means that contribute to the ends in the nodes on the higher layers.

Besides the *pragmatic*, how-to-act, means-end aspect just explained, we may distinguish at least two other, largely independent, dimensions along the edges of the objectives tree. The first is a *semantic* one. From top to bottom, the notions in the nodes of the tree vary from abstract to concrete, and the lower-level nodes operationalize the higher-level ones. From this semantic perspective, the tree explicates what the higher-level objectives mean in relation to the artifact and its context. For instance, "safe" in relation to a transport system may be operationalized among other things with "low number of deaths." To show that

the pragmatic aspect of "safe" in the tree differs from the semantic one, we need only realize the following. "High speed of medical response to accidents" serves the purpose of "low number of deaths," which serves the purpose of being a "safer transport system." We can hardly claim however that high speed of medical response to accidents makes the transport system safer. Apparently, the pragmatic aspects along the edges of the tree are transitive where the semantic perspective sometimes lacks this property. Besides pragmatics and semantics, we may also distinguish a *value* dimension along the branches of the tree. Every node, but the highest one, has instrumental value for connected nodes higher in the tree, and the highest node has only an intrinsic value. Normally, the weight of the node values varies with their level in the tree – the higher, the more important. The lowest one may even have a negative value, so that we come to situations where "the end justifies the means." Although not completely unrelated, the means-end and the value dimensions in the tree are not identical and need to be distinguished.

An interesting proposal to systematize and explicate modeling for values can be drawn from Van de Poel (2013). Van de Poel combines the three dimensions of the designers' objectives tree to operationalize the abstract values at the top of the tree using the values of the leaves at the bottom. Since models are artifacts, his approach is also relevant for model builders. To realize abstract values in a design, Van de Poel introduces *values hierarchies*, which consist of three basic layers: the *abstract values* relevant for the artifact reside at the top layer; the middle layer consists of all *general norms*, which are considered as "prescriptions for, and restrictions on, action" (p. 258); and the bottom layer consists of the design requirements. Van de Poel considers two criteria for the top-down operationalization of abstract values based on norms: "the norm should count as an appropriate response to the value" and "the norm, or set of norms, is sufficient to properly respond to or engage with the value." In a second step, these norms are specified with the aid of design specifications. This step may concern the explication of the goal, the context, and the action or the means.

Bottom-up values hierarchies are built up from *for-the-sake-of relations*. Design requirements serve the purposes of certain norms, which on their turn are built in for the sake of the final and abstract norms. Van de Poel discusses the example of chicken husbandry systems. There, general value animal welfare is served by the general norms: presence of laying nests, litter, perches, and enough living space. On its turn, these norms are realized by design requirements such as at least 450 $cm^2$ floor area per hen, 10 cm of feeding trough per bird, and floor-slope of maximal 14 %.

As we saw in the first section, engineers already design for general values such as safety, sustainability, and even privacy. The point is, however, the following. If modelers (and engineers) were to introduce values hierarchies as an instrument to realize these values explicitly, the way the model serves certain values and avoids negative ones would be much more explicit, debatable, and open for corrections and improvements. Surely, values hierarchies do not solve all value conflicts in modeling and design, but at least they explicate and systematize the value judgments involved in modeling and design. By doing so, one renders these judgments

transparent for discussion in internal and external debates.[17] This transparency of the values implied by models and artifacts is a necessary condition for launching new models and artifacts into a civilized and democratic society.

Although this chapter emphasizes the parallels between the methods applied by designers of technical artifacts and modelers, who model for societal values, we should not forget an important relative difference between the two. Designers deciding about pure technological matters have a wealth of scientific and engineering literature to consult and experiments to carry out; they have much more objective or intersubjective knowledge about the world to fall back on than modelers, managers, and politicians who should decide about the societal and environmental effects of a model. Take, for example, the design of some alloy steel with a certain strength, corrosion resistance, and extreme temperature stability for nuclear reactors. The designer who decides how to design steel with the required specifications can fall back on material science and can carry out experiments. Her or his design will be based on a wealth of intersubjective knowledge and experience. This decision process is normative but backed up more strongly by scientific knowledge than decisions concerning the societal, political, and environmental impacts of steel production. Questions such as should nuclear reactors be made at all and, if so, which rare earth metals should be used and in which countries we can find these metals are less straightforwardly backed up by science. In one word, societal and environmental values are backed up with far less objective and generally accepted knowledge than scientific and technological ones. The same holds *mutatis mutandis* for modelers and their models.

Because knowledge about technical values differs from knowledge about societal and environmental values, the question arises of who should decide about the latter ones. Engineering modelers seem the most appropriate for making strict technical decisions about modeling and design. Are they however also the ones that should take the decisions regarding societal and environmental issues? Since some designers have concluded that this question should be answered negatively, they initiated the democratization movements mentioned above. *Mutatis mutandis,* the same could be said about modeling. The values-hierarchy method can be carried out by a modeler or a modeling team. Nevertheless, it is to be preferred and less paternalistic or even more democratic, and even more just, when the value decisions are taken by all stakeholders involved.

## Documentation About the Values

In the previous section, we discussed an instrument helping to take into account values as explicitly, transparently, and systematically as possible. The GIS and Patriot examples show that if modelers want to evade societal and environmental

[17]These are two ends that also inspired the cautious admitters' position of Le Menestrel and Van Wassenhove discussed in the section "Current Ideas About the Value-Ladenness of Models."

accidents, their task does not end with only applying this or similar instruments. They should stimulate and further the value debate among colleagues, users, and other stakeholders. To avoid accidents, modelers should also provide extensive technical documentation and user manuals. In addition, one could even argue that parallel to the designers of artifacts, they should provide aftercare for their creations and should evaluate how their products function in the real world.

In most general terms, following a design methodology such as that proposed in the fifth chapter of Whitaker and Mancini (2012) would enable modelers to document value-sensitive decisions made during the design of the model in the same sense. Whitaker and Mancini state that the documentation production in system engineering follows the four-design-phase cycle mentioned before. After having discussed the iterative nature of design processes, they claim:

> "At each stage in the process, a decision is made whether to accept, make changes, or return to an earlier stage of the process and produce new documentation. The result of this activity is documentation that fully describes all system elements and that can be used to develop and produce the elements of the system." Whitaker and Mancini (2012, p. 69)

Identifying and following the same stages in the modeling process, modelers could hold track of the decisions they make about envisaged instrumental and even derived values involved. Doing so, they could provide the model users and other stakeholders with technical documentation about their analyses and decisions regarding these values embedded and implied by their creations.

As we have done previously, here again we should distinguish between modeling situations in which the modeling and design teams are close or even identical and those in which the distance between the two is much larger. In the first case, the modelers should be at least as much engaged in the design development phases of the artifact as in those of the model. To produce adequate value documentation, the modelers should be acquainted with the value assessments of the designers and with their ideas about proper and improper use of their artifact in practice. For the modelers, then, the emphasis is on the derivative values of their models; the Patriot case provides a telling example. In the second case, characterized by a large distance between the modeler and user, this emphasis is more on the instrumental or goal-related values of the models. Modelers should therefore support and document the value assessments following their own modeling methodology – the GIS case provides a good example. The first situation, with a small distance between modelers and users, compares with what in the technical-documentation literature has been called *self-documentation*, characterized by small enterprises and close collaboration (Baumgartner and Baun 2005, p. 2384); the second one, which relates to a large enterprise, more complex tasks and a large distance between the collaborators, requires *database documentation* (idem, p. 2385).

Besides the general documentation discussions, *checklists* of items that such documentation should cover are helpful. Below, I attempt to set up such a list without claiming its entries are necessary and the list is sufficient or complete. This first attempt should be read as an invitation to modelers and colleagues to discuss and elaborate, such that we come to a more mature list ratified by modelers and

analysts studying modeling practices. Setting up the list, I envisaged the distance between the modelers and the users to be considerable, and following the list one is likely to end with a value description that is more like a database than self-documentation. The list features mainly instrumental values and general derivative ones because concrete ones depend too much on the details of the artifact and its context. The issues that follow emanate mainly from definitional characteristics, engineering and model building practices, and societal and environmental values.

To my mind the values within the model documentation should at least cover the following items (their origins are mentioned between brackets):

1. Clear indications about the purpose of the model (i.e., what it is made for and what is its function) and a description of its proper use (i.e., how it should be used according to its makers to achieve the goal of using it) (purposes)
2. The list of the model's design specifications and clear descriptions of how, and in which context, the model should be applied within its window of application (purpose)
3. Indications about the model's limitations, its abstractions, and its assumptions (approximation and representation)
4. Clear indications about the model's technical properties and behavior such as its robustness and efficacy, and information about how the model was verified (technical values)
5. Clear indications about the model's accuracy and about its validation (epistemic values)
6. A clear description of the tensions between various values in the model (and the resulting artifact) and what choices have been made to cope with them (transparency and communication)
7. Indications about how the model (with or without the supported artifact) copes with the engineering values such as safety, risk, reliability, security, effectiveness, and the costs, within and even outside the specs
8. Statements about how the model in isolation or in combination with the intended artifact takes into account societal and environmental values such as the quality of individual life, of social life, and of the environment
9. Descriptions of how the models have applied defensive design methodologies to make the model (or model artifact combination) foolproof and thus prospectively prevent possible model accidents, which may be due to all kinds of misuse such as application outside the specs or use for other purposes than intended
10. Plans about how the introduction of the model will be monitored, whether or not in combination with its artifact, and possibly adapted, adjusted, or improved when it turns out that its use implicates negative societal consequences (aftercare)

Entry 1 is needed to learn about the proper use of the model and to delimit the scope of the model's application. Some models are developed for general applications, while others are optimized only under specific conditions. Ad 2. The explicit

list of the model's design specifications and, if very complex, a summary of this list enable close and distant users to learn about the details of the model's scope of application. Entry 3 should be covered because as models are approximate representations, they need to leave out many issues of the represented target system. If these left-out issues are not acknowledged in the documentation, model users might have false beliefs about the abilities of a model. The model's abstractions are closely related to its assumptions. Every model incorporates a variety of assumptions that have fundamental impact on the performance of a model. Items 4 and 5 are necessary for a responsible launch of the model in society.

Ad 6. For transparency's sake, all designs have to cope with tensions between values, and the design of models makes no exception. The value documentation should explicate which tensions the modelers and designers have considered and how they managed these tensions using which arguments. For instance, which compromise between privacy and efficacy has been chosen? How are errors of type I and type II balanced and for what reasons? Ad 7. Since technical and engineering values may conflict with societal and environment ones, modelers should explicate all of them to chart these tensions and their choices. Entry 8 requires modelers and designers to explicate what they did to anticipate all relevant societal and environmental issues. Ad 9. In principle, defensive design comes down to anticipating all ways in which an artifact can be misused and blocking misuse or reducing the damage by adequate design. Of course, according to Murphy's law, complete foolproof models and artifacts do not exist. Douglas Adams wrote succinctly: "common mistake that people make when trying to design something completely foolproof is to underestimate the ingenuity of complete fools" (Adams 2009, p. 113). Finally, in item 10, modelers should explain their plans for aftercare.

The checklist illustrates that the design, development, and the introduction of complicated models and artifacts in society are a complicated combination of retro- and prospective processes. Technical (4) and epistemic properties (5) of a model are often only assessed in a backward-looking way. Planning the aftercare (10) and assessing the possible impact of the model or model artifact combination (9) are clearly forward looking. Other issues are combinations of the two. Establishing the engineering values (7), for instance, is based on past experience, but the model's performance regarding these values will never be completely certain and should be monitored when being used in practice. Even the models' abstractions (3), their specifications (2), and even (1) their exact proper uses are not for always fixed in advance. Research in engineering practices such as of Downey (1998) and Vinck (2003) but especially that of Bucciarelli (1994) clearly shows that in real life engineering design is a social practice. The lists of specifications, the exact purpose of the design, and even the working principles might drastically change during the design process and for social-technical systems even after the design has been launched into society. This adaptation of the specs during and even after the design process makes Bucciarelli and Kroes (2014) claim that instrumental rationality falls short of describing the engineering design in practice. The forward-looking aspect of introducing models and artifact in society makes it partly an *open* process.

The openness of the development and introduction of technical models, standalone or underlying a technical artifact, is the reason for the traditional deniers' argument of value-free models to fail. Model development and introduction is a dynamical and open process and is an interaction between many stakeholders. The model's specs, its proper use, and the consequences of improper use are partly un- and underdetermined, and many of these consequences only become clear once the artifact is launched into society. Because modelers are the most acquainted with the model's behavior within and outside the specs, they are at least accountable for determining the specs, the choice of the borderlines between the proper and improper use, and other value-related issues. The openness of launching artifacts or models into society obliges their creators to participate in a prospective process of investigating how their model might cause damage to persons, society, or the environment. Considering, for example, the GIS and the Patriot cases, modelers and designers should take the lessons at heart of Merton and his followers about unforeseen or unanticipated circumstances. Ample and comprehensive model documentation helps to prevent the modeler to become blameworthy regarding possible damage inflicted by her or his artifact. Moreover, and perhaps even more importantly, it helps to make the discussion about the target values of the modeling process much more explicit and transparent; by doing so, it justifies and democratizes the processes of modeling and engineering design.

## Summary and Conclusions

This chapter has been mainly about the instrumental and derivative values of models in engineering design. For the common technical and epistemic modeling values such as safety and validity, it referred to the standard literature. We focused mainly on issues regarding how social and environment values emerge unnoticed during the process of developing and applying the model. Because launching models and their affordances into society is an open process, we argued that modelers have more responsibility than just the behavior within the specs. Next, we showed how various forms of indeterminacy, underdeterminacy, complexity, and lack of communication about proper use may have (often unnoticed) value-laden consequences in practice. An important way to model for values is then to take models explicitly to be special kinds of artifacts and apply various design methodologies. This way of interpreting the modeling job enables the modeler to apply methods and techniques from design and to view a modeling problem as a multiple-criteria problem, which is in need of an explicit list of design specifications including value-related issues. To find all these values, modelers may apply forms of stakeholder analysis and participatory design. Additionally, they can apply hierarchical means-end trees to explicate and operationalize the values and their mutual tensions involved in the modeling job, supporting their internal and external discussions. Finally, the model-as-artifact perspective helps modelers to sustain this discussion by producing technical documentation and user guides during the various phases of the modeling (design) process. The chapter ended with a checklist of

issues, which the documentation should encompass if a modeling team wants to make a start with taking modeling for values seriously. May this chapter be a first step toward more comprehensive methods and lists for managing societal values in modeling engineering design.

Modelers should realize that they can, and often should, model for certain values, not the least, because they are accountable for negative (and positive) societal implications of their creations. Because of this, they should not only take care of the functional, technical, and engineering values of their creations. They should also proactively spot unanticipated societal implications of their contrivances. To systematize the instrumental and derivative values and their tensions, they can apply values hierarchies to manage, realize, and document these values in their work. In doing so, modelers would render the value-ladenness of their work more transparent and would contribute substantially to the internal and public debate about the social and environmental consequences of their models.

**Acknowledgment** This chapter draws on and elaborates Zwart et al. (2013) and Diekmann Zwart (2013). Moreover, it presents part of Van de Poel (2013) as starting point for the operationalization of societal values in engineering design. Finally, the author wants to thank Sven Diekmann and the editors of the present volume for their comments on the outline and contents of this chapter.

## Cross-References


- ▶ Conflicting Values in Design for Values
- ▶ Design for the Values of Accountability and Transparency
- ▶ Design for the Values of Democracy and Justice
- ▶ Design for Values in Engineering
- ▶ Design Methods in Design for Values
- ▶ Design for Values and the Definition, Specification, and Operationalization of Values
- ▶ Participatory Design and Design for Values


## References


ABET, Accreditation Board for Engineering and Technology, Inc (1988) Annual report for the year ending September 30, 1998, New York

Adams D (2009) Mostly harmless. Pan Macmillan, London

Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. Syst Dyn Rev 12(3):183–210. doi:10.1002/(SICI)1099-1727(199623)12:3<183::AID-SDR103>3.0.CO;2-4

Baumgartner F, Baun TM (2005) Engineering documentation. In: Whitaker JC (ed) The electronics handbook, 2nd edn. CRC Press, Boca Raton

Bijker WE (1995) Democratisering van de technologische cultuur. Schrijen-Lippertz, Voerendaal

Blair M, Obenski S, Bridickas P (1992) GAO/IMTEC-92-26 Patriot missile software problem. Retrieved from http://www.fas.org/spp/starwars/gao/im92026.htm

Bucciarelli LL (1994) Designing engineers. MIT Press, Cambridge, London

Bucciarelli L, Kroes P (2014) Values in engineering. In: Soler L, Zwart S, Lynch M, Israel-Jost V (eds) Science after the practice turn in the philosophy, history, and social studies of science. Routledge, NewYork/Londen, pp 188–199
Buchanan R (1992) Wicked problems in design thinking. Des Issues 8(2):5–21. doi:10.2307/1511637
Cranor CF (1990) Some moral issues in risk assessment. Ethics 101(1):123–143. doi:10.2307/2381895
Cross N (2008) Engineering design methods: strategies for product design. Wiley, Chichester/Hoboken
Diekmann S, Zwart SD (2013) Modeling for fairness: a rawlsian approach. Stud Hist Philos Sci A 46:46–53
Douglas H (2000) Inductive risk and values in science. Philos Sci 67(4):559–579
Douglas H (2007) Rejecting the ideal of value free science. In: Kincaid H et al (ed) Value-free science?, vol 1. Oxford University Press, New York, pp 120–141
Downey GL (1998) The machine in me. An anthropologist sits among computer engineers. Routledge, New York/London
Feenberg A et al (2006) "Replies to critics", democratizing technology: Andrew Feenberg's critical theory of technology. In: Veak TJ (ed) Democratizing technology: building on Andrew Feenberg's critical theory of technology. State University of New York Press, Albany, pp 175–210
Fleischmann KR, Wallace WA (2005) A covenant with transparency: opening the black box of models. Commun ACM 48(5):93–97. doi:10.1145/1060710.1060715
Fleischmann KR, Wallace WA (2009) Ensuring transparency in computational modeling. Commun ACM 52(3):131–134. doi:10.1145/1467247.1467278
Frankena WK (1973) Ethics. Prentice-Hall, Englewood Cliffs
Frigg R, Hartmann S (2012) Models in science. In: Zalta EN (ed) The stanford encyclopedia of philosophy (Fall 2012 edition). The Metaphysics Research Lab Stanford, CA 94305-4115 Stanford. http://plato.stanford.edu/archives/fall2012/entries/models-science/
Gelfert A (2009) Rigorous results, cross-model justification, and the transfer of empirical warrant: the case of many-body models in physics. Synthese 169(3):497–519. doi:10.1007/s11229-008-9431-6
Geraci A (1991) IEEE Standard Computer Dictionary: Compilation of IEEE Standard Computer Glossaries. (Contributions by F. Katki, L. McMonegal, B. Meyer, J. Lane, P. Wilson, J. Radatz, . . . F. Springsteel). Piscataway, NJ, USA: IEEE Press
Gibson JJ (1986) The ecological approach to visual perception. Lawrence Erlbaum, Hillsdale
Haimes YY (2005) Risk modeling, assessment, and management, vol 40. Wiley, Hoboken
Hubka V, Eder WE (1988) Theory of technical systems; a total concept theory for engineering design. Springer, Berlin
ISO (2006) ISO 11442:2006(E) Technical product documentation – document management. International Organization for Standardization, Geneva
Jenkins DG, McCauley LA (2006) GIS, SINKS, FILL, and disappearing wetlands: unintended consequences in algorithm development and use. In: Proceedings of the 2006 ACM symposium on applied computing. ACM, New York, pp 277–282. doi:10.1145/1141277.1141342
Jones JC (1992) Design methods. Wiley, New York
Kijowski DJ, Dankowicz H, & Loui MC (2013) Observations on the Responsible Development and Use of Computational Models and Simulations. Science and Engineering Ethics, 19 (1):63–81. doi:10.1007/s11948-011-9291-1
Klaasen I (2005) Modelling reality. In: Jong TMD, Voordt VD (eds) Ways to study and research urban, architectural and technical design. IOS Press/Delft University Press, Delft, pp 181–188
Kleijnen JPC (2001) Ethical issues in modeling: some reflections. Eur J Oper Res 130(1):223–230. doi:10.1016/S0377-2217(00)00024-2
Le Menestrel M, Van Wassenhove LN (2004) Ethics outside, within, or beyond OR models? Eur J Oper Res 153(2):477–484. doi:10.1016/S0377-2217(03)00168-1
Mannan S (2005) Lee's loss prevention in the process industries: hazard identification, assessment, and control. Elsevier Butterworth-Heinemann, Burlington

McNelis PD (1994) Rhetoric and rigor in macroeconomic models. In: Wallace WA (ed) Ethics in modeling. Pergamon, Oxford/Tarrytown, pp 75–102
Merton RK (1936) The unanticipated consequences of purposive social action. Am Sociol Rev 1 (6):894–904. doi:10.2307/2084615
Morgan MS, Morrison M (1999) Models as mediators: perspectives on natural and social science. Cambridge University Press, Cambridge
Pahl G, Beitz W (1984) Engineering design; a systematic approach. Design Council, London
Pugh S (1990) Total design; integrated methods for successful product engineering. Addison Wesley, Wokingham
Rittel HWJ, Webber MM (1973) Dilemmas in a general theory of planning. Policy Sci 4 (2):155–169. doi:10.1007/BF01405730
Roozenburg NFM, Eekels J (1995) Product design: fundamentals and methods. Wiley, Chichester/New York
Rykiel EJ (1996) Testing ecological models: the meaning of validation. Ecol Model 90 (3):229–244. doi:10.1016/0304-3800(95)00152-2
Sargent RG (2005) Verification and validation of simulation models. In: Proceedings of the 37th conference on winter simulation, Orlando pp 130–143
Schuler D, Namioka A (1993) Participatory design: principles and practices. Lawrence Erlbaum, Hillsdale
Selby RG, Vecchio FJ, Collins MP (1997) The failure of an offshore platform. Concrete Int 19 (8):28–35
Shelley C (2011) Fairness in technological design. Sci Eng Ethics 18(4):663–680. doi:10.1007/s11948-011-9259-1
Shruti K, Loui M (2008) Ethical issues in computational modeling and simulation. Cincinnati
Siegel, Adam B (2004) "Honest Performance Analysis: a not-always met requirement". Defense Acquisition Review Journal. Defense Acquisition University Press. January–April, p.101–106
Simon HA (1973) The structure of ill structured problems. Artif Intell 4(3–4):181–201. doi:10.1016/0004-3702(73)90011-8
van de Poel IR (2009) Values in engineering design. In: Meijers AA (ed) Philosophy of technology and engineering sciences, vol 9. Elsevier/North Holland, Amsterdam/London/Boston, pp 973–1006
Van de Poel I (2011) The relation between forward-looking and backward-looking responsibility. In: Vincent NA, van de Poel I, Hoven J (eds) Moral responsibility. Springer Netherlands, Dordrecht, pp 37–52
van de Poel IR (2013) Translating values into design requirements. In: Michelfelder DP, McCarthy N, Goldberg DE (eds) Philosophy and engineering: reflections on practice, principles and process. Springer, Dordrecht/Netherlands, pp 253–266
van de Poel IR, Royakkers L (2011) Ethics, technology, and engineering: an introduction. Wiley-Blackwell, Malden
Vinck D (ed) (2003) Everyday engineering. Ethnography of design and innovation. MIT Press, Cambridge
Walker WE (1994) Responsible policy making. In: Wallace WA (ed) Ethics in modeling. Pergamon, Oxford/Tarrytown, pp 226–241
Walker WE (2009) Does the best practice of rational-style model-based policy analysis already include ethical considerations? Omega 37(6):1051–1062. doi:10.1016/j.omega.2008.12.006
Whitaker JC, Mancini RK (2012) Technical documentation and process. CRC Press, Boca Raton
Woodhouse E, Patton JW (2004) Design by society: science and technology studies and the social shaping of design1. Des Issues 20(3):1–12. doi:10.1162/0747936041423262
Zeigler BP, Praehofer H, Kim TG (2000) Theory of modeling and simulation, 2nd edn. Academic, San Diego
Zwart SD, Jacobs J, van de Poel I (2013) Values in engineering models: social ramifications of modeling in engineering design. Eng Stud 5(2):93–116

# Part III

# Values

# Design for the Values of Accountability and Transparency

Joris Hulstijn and Brigitte Burgemeestre

## Contents



**Abstract**

If an organization is to be held accountable for its actions, the public need to know what happened. Organizations must therefore "open up" and provide evidence of performance to stakeholders, such as principals who have delegated tasks, employees, suppliers or clients, or regulators overseeing compliance. The social

J. Hulstijn (✉)
Delft University of Technology, Delft, The Netherlands
e-mail: j.hulstijn@tudelft.nl

B. Burgemeestre
Delft University of Technology, Delft, The Netherlands

Pandar, Amsterdam, The Netherlands
e-mail: brigitte@burgemeestre.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_12

values of transparency – the tendency to be open in communication – and accountability – providing evidence of past actions – are crucial in this respect.

More and more aspects of the internal control systems, policies, and procedures to gather evidence of organizational performance are implemented by information systems. Business processes are executed and supported by software applications, which therefore effectively shape the behavior of the organization. Such applications are designed, unlike practices which generally grow. Therefore, it makes sense to take the core values of accountability and transparency explicitly into account during system development.

In this chapter we provide an account of the way in which transparency and accountability can be built into the design of business processes, internal controls, and specifically the software applications to support them. We propose to make trade-offs concerning core values explicit, using an approach called value-based argumentation. The approach is illustrated by a case study of the cooperation between providers of accounting software and the Dutch Tax and Customs Authority to develop a certificate, in order to improve the reliability of accounting software. Widespread adoption of the certificate is expected to stimulate accountability and indeed transparency in the retail sector.

Although the approach is developed and tested for designing software, the idea that trade-offs concerning core values can be made explicit by means of a critical dialogue is generic. We believe that any engineering discipline, like civil engineering, water management, or cyber security, could benefit from such a systematic approach to debating core values.

**Keywords**
Value sensitive design • Accountability • Transparency

## Introduction

From the beginning of the 1990s, there has been an audit explosion: formal audit and evaluation mechanisms have become common practice in a wide variety of contexts (Power 1997). Not only in the financial domain where it originates but also in other domains like healthcare or education, professionals are now expected to be accountable to the general public by means of extensive reporting schemes. The advance of auditing is closely related to the increasing importance of the social values *accountability*, providing evidence to justify past actions to others, and *transparency* – the tendency to be open in communication. In this chapter we would like to investigate to what extent these values affect systems design. Is it possible to design a system taking accountability and transparency into account? In brief, can we design for accountability or for transparency?

Accountability and transparency are properties of a person or organization. When used to describe a system, their meaning is derived. They are iconic notions, with positive connotations. For example, accountability is used as a kind of synonym to good governance (Dubnick 2003). Transparency of government has

become a goal in its own right, witness, for example, in the movement for open data (Zuiderwijk and Janssen 2014). Taken in this way, these notions are used as evaluative, not as analytical, concepts. Relative to what or whom can we evaluate these notions? Clearly their meaning depends on the context, on a relationship with others: "Accountability can be defined as a social relationship in which an actor feels an obligation to explain and to justify his or her conduct to some significant other" (Day and Klein 1987). According to Bovens (2005, 2007), an accountability relationship contains a number of components: the actor can be a person or agency. Then there is some significant other, who can be a person or agency but can also be a more abstract entity, such as God or "the general public." Bovens calls this the forum. The relationship develops in three stages. First, the actor must feel obliged to inform the forum about its conduct, including justifications in case of failure. The obligation may be both formal, i.e., required by law or by contract, or informal and self-imposed, for instance, because the actor is dependent on the forum. Second, the information may be a reason for the forum to interrogate the actor, ask for explanations, and debate the adequacy of the conduct. Third, the forum passes judgment on the actor's conduct. A negative judgment often leads to some kind of sanction; again this can be both formal and informal. This means that an accountability relation should provide room for discussion: it is not a one-way stream of reports but rather a dialogue. The dialogue is facilitated by a transparent organization and by an inquisitive forum challenging the outcomes. Moreover, accountability is not without consequences. There is a judgment that depends on it.

We can also analyze accountability as the counterpart of responsibility (Van De Poel 2011). When I am responsible for my actions now, I may be held accountable later. I need to collect evidence, so that I can justify my decisions. One could say that this focus on evidence collecting has turned a moral topic into a rather more administrative or technical one.

In this chapter we will focus on a specific application of accountability in which evidence plays an important role, namely, regulatory compliance. Companies are accountable for their conduct to the general public, represented by a regulator (e.g., environmental inspection agency, tax administration). In corporate regulation, many laws nowadays involve some form of self-regulation or co-regulation (Ayres and Braithwaite 1992; Black 2002). Regulators increasingly rely on the efforts of the companies being regulated. Companies must determine how the regulations apply to their business, set up a system of controls to ensure compliance, monitor the effectiveness of these controls to establish evidence, and subsequently provide accountability reports. To ensure reliability, stakeholders demand certain guarantees in the way evidence is being generated. These measures are called internal controls (Coso 1992). Consider, for example, segregation of duties, reliable cash registers, automated checks, access control, or logging and monitoring.

Crucially, such internal control measures are being designed. An adequate design of the internal controls is a prerequisite for this kind of accountability. But it is not enough. Evidence is being generated in a corporate environment. People are often in the position to circumvent the controls or manipulate their outcomes

(Merchant 1998). Whether they choose to do so depends on their values and beliefs, which are partly determined by the corporate culture (Hofstede et al. 1990). In particular the value of transparency is important here: do we really want to reveal these facts about our conduct? Corporate culture cannot be designed; it can only be stimulated. For instance, transparency can be facilitated by systems that make it easier rather than harder to access and share information.

When an auditor or inspector is asked to verify information disclosed by a company, he or she must rely on the controls built into the procedures, processes, and information systems. As a consequence, there is an increased need for methods to support the systematic design of control measures and secure information systems (Breu et al. 2008). Such methods are developed in the field of requirements engineering. There is a whole set of requirements engineering techniques specifically targeted to make systems secure (Dubois and Mouratidis 2010). However, these methods focus on the design of controls from a risk-based point of view: which threats and vulnerabilities are addressed by which measures? They pay little attention to the facilitation of the underlying values. In fact, security may even be harmful to transparency. Security measures often make it harder to share information.

In the area of information security, the notion of accountability has often been confused with traceability: the property of being able to trace all actions to an identifiable person in a specific role. However, traceability is neither a necessary nor sufficient condition to establish accountability (Chopra and Singh 2014). So accountability and transparency need to be considered in their own right. That suggests the following research question:

How can we make sure that the values of accountability and transparency are built into the design of information systems?

There are design methodologies that aim to incorporate nonfunctional requirements or values like security (Fabian et al. 2010), accountability (Eriksén 2002), or even transparency (Leite and Cappelli 2010) into the development process. However, these values compete with other values, such as profitability, secrecy, and human weaknesses such as shame. How can such conflicts be resolved? Following the recommendation of Bovens (2007), we take the notion of a dialogue very seriously. The general idea is that design is a matter of trade-offs (Simon 1996). Core values, like transparency and accountability, need to be defended against other values and tendencies, such as efficiency, profitability, tradition, or secrecy. That means that design choices are justified in a process of deliberation and argumentation among a forum of stakeholders. Our working hypothesis is that such a critical dialogue can be facilitated by a technique called *value-based argumentation* (Atkinson and Bench-Capon 2007; Atkinson et al. 2006; Burgemeestre et al. 2011, 2013). One stakeholder proposes a design. Other stakeholders can challenge the underlying assumptions and ask for clarification. In this way, more information about the design is revealed. Ultimately, the forum passes judgment. The resulting scrutiny should provide better arguments and subsequently improve the quality and transparency of the design process itself.

In order to illustrate the dialogue approach and demonstrate its adequacy in handling design trade-offs, we describe a case study. The case study is about a

long-term cooperation between a number of software providers and the Netherlands Tax Administration. These partners developed a set of standards to promote reliability of accounting software, make sure that crucial controls are embedded in the software, and thus ensure reliability of financial recording. This is crucial for tax audits, but it also affects trustworthiness of businesses in general. By initiating a standard, the government is trying to stimulate transparency and accountability in the entire business community.

Although we focus here on the design of accounting software, the use of value-based argumentation is generic. Values are crucially related to the identity of people or groups (Perelman 1980). They reveal something about who you are or where you have come from. This explains why debates about values run deep. We believe that other engineering disciplines, which must deal with challenges like safety, sustainability, or environmental hazards, could also benefit from such a systematic approach to debating core values.

The remainder of the chapter is organized as follows: the following section explains the notions of accountability and transparency in more detail. Then we develop the idea that one can in fact design for accountability, and possibly also for transparency. After that the case study is discussed. The chapter ends with conclusions and directions for further research.

## Accountability and Transparency

### Auditing, Agency Theory, and Internal Controls

As we stated in the introduction, we focus on accountability and transparency in the context of regulatory compliance. Companies collect evidence of behavior and produce reports. To verify these reports, auditors or inspectors are called in. Therefore, auditing theory is relevant here; see textbooks like Knechel et al. (2007). "Auditing is the systematic process of objectively obtaining and evaluating evidence regarding assertions about economic activities and events to ascertain the degree of correspondence between the assertions and established criteria, and communicate the results to interested users" (American Accounting Association 1972). Roughly, auditing is testing to a norm (Fig. 1). What is being tested is a statement or assertion made by management about some object, for instance, the accuracy and completeness of financial results, the reliability of a computer system, or the compliance of a process. The statement is tested against evidence, which must be independently collected. The testing takes place according to norms or standards. In accounting, standards like the Generally Accepted Accounting Principles (GAAP) are well established, but in other fields, like regulatory compliance, the standards often have to be developed, as the original laws are written as open norms. Open norms are relatively general legal principles or goals to be achieved; they still have to be made specific for the situation in which they are applied (Korobkin 2000; Westerman 2009). Often, determining the boundaries leads to a kind of dialogue among stakeholders. Black (2002) calls such dialogues regulatory conversations.

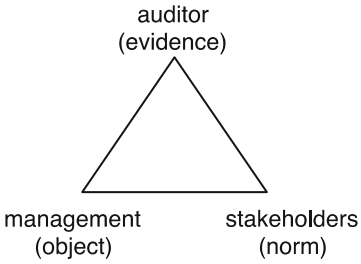



**Fig. 1** Typical audit setting

Much recent thinking about accountability and compliance monitoring has come to be dominated by a rather mechanical logic of auditability (Power 2009). This way of thinking is characterized by a bureaucratic demand for evidence and by reference models like COSO and COBIT that try to put reality into a rational mold of control objectives and measures to mitigate risks. The plan-do-check-act loop (Deming 1986) that was originally developed for improving quality in the automobile industry is widely adopted to make organizations learn and improve their risk management efforts (Power 2007). This Deming cycle is essentially a feedback-control loop, which presupposes that an organization is run as a machine, with levers and dials. Although corporate reality rarely fits these molds, accountability has sometimes degenerated into a box-ticking affair.

In the typical audit relation, we identify three actors: management, stakeholders, and auditors (Fig. 1). Management is accountable for its actions to the owners or shareholders of a company, who want to see a return on investment. It is also accountable to other members of society, like employees, suppliers, or others, who are dependent on the company. Every year management prepares financial statements about the results of the company: profit and loss, assets and expectations, and compliance. Auditors verify these statements and provide assurance as to their reliability. In this case, the accountability derives from a delegation of tasks: shareholders have delegated executive tasks to management. The resulting loss of control is remedied by accountability reporting.

This type of accountability relationship is typically addressed by agency theory (Eisenhardt 1989). One party, the principal, delegates work to another party, the agent. The agent must give an account of his or her actions, because the principal is distant and unable to verify or control the agent's actions directly (Flint 1988). In addition, the agent's incentives may conflict with those of the principal, so the agent may withhold evidence of executing the task. The resulting information asymmetry is one of the focal points of agency theory. The principal runs a risk, for two reasons: (i) she is dependent on the agent for executing the action, but does not have a way of directly controlling the agent, and (ii) she is dependent on the agent for providing evidence of execution. To overcome these risks, the principal will typically demand guarantees in the way information is being generated: internal controls.

Similar examples can also be found in other domains. Consider forms of self-regulation (Rees 1988). A company states that its products adhere to health and

safety regulations. To keep its license, the company must report on performance indicators to demonstrate that it is meeting its objectives. Inspectors regularly evaluate these claims. Here, the general public acts as principal: it has delegated meeting part of its concerns (health and safety) to the company itself. The inspector must provide assurance. This form of self-regulation creates information risks similar to (i) and (ii) above. Such regulatory supervision schemes are discussed in Burgemeestre, et al. (2011).

In agency theory, information systems are seen as a means both to control behavior and to inform the principal about what the agent is doing (Eisenhardt 1989). Control measures are often embedded in information systems. Consider, for example, application controls, which are built into financial applications and automatically prevent common mistakes in transactions, like entering a negative price or a nonexistent product code. Consider also controls built into a point-of-sales system at a supermarket, which make it hard or impossible to delete the initial recordings of a sales event. Or consider the automatic maintenance of segregation of duties, which ought to make sure that some transactions or data files can only be accessed or manipulated by people in specific roles. All of these measures are implemented in or at least facilitated by information systems. Important facilitating functions, needed for reliable information provisioning and control, are baseline security, logging and monitoring, access control, and authorizations management (Romney and Steinbart 2006).

## Accountability

Even though accountability has become a main issue in today's audit society (Power 1997) and information systems are crucial for the collection of evidence, the topic has received limited attention in the computer science community. When accountability is discussed in computer science, it often focuses on the human accountability for some computer program and not on the accountability achieved or supported by the software itself. For example, it is debated whether software agents or their human owners can be held morally accountable for their actions (Friedman et al. 1992; Heckman and Roetter 1999). Can programmers be held accountable for problems that occur due to errors in the software they produced (Johnson and Mulvey 1995; Nissenbaum 1994)?

One of the few works that does describe how accountability can be incorporated into an information system is the work of Friedman et al. (2006) on *value sensitive design*. They define accountability as the set of properties of an information system that ensure that the actions of a person, group of people, or institution may be traced uniquely to the person, people, or institution (Friedman et al. 2006). So under this view, accountability is ensured by an audit trail. In the context of regulatory compliance, also Breaux and Anton (2008) consider a software system to be accountable if "for every permissible and non-permissible behaviour, there is a clear line of traceability from the exhibited behaviour to the software artefacts that contribute to this behaviour and the regulations that govern this behaviour" (p. 12).

In their view, a system is accountable when it is able to demonstrate which regulatory rules apply to which transaction and thus produce a corresponding audit trail.

However, Chopra and Singh (2014) argue that although such traceability is an important mechanism for holding someone accountable, it is neither necessary nor sufficient. First, traceability of actions is not always necessary. One can also use the outcomes of a process, rather than the way it was carried out, to hold someone accountable. Compare the difference between outcome control and behavioral control (Eisenhardt 1985). Second, traceability is not enough. What is needed in addition is a mechanism of holding the agent accountable: someone must evaluate the audit trail and confront the agent with possible deviations. This is precisely the function of the forum (Bovens 2007).

Who plays the role of the forum? In trade relationships, we often find actors with countervailing interests (e.g., buyer and seller) who can hold each other accountable. This is also why segregation of duties is considered so important: it creates independent sources of evidence, which can be used for cross verification. In bureaucracies, often an artificial opposition is created, for instance, between the front office (help client) and the back office (assess conformance to policies). Also the installation of a dedicated risk management function, separated from the business, can be seen in this respect. When effective, such a risk function should provide a counterforce against the tendency of the business to take on too many risks; see, e.g., (Coso 2004; Power et al. 2013). The resulting critical dialogue between the business and risk function should lead to lower risks and motivated controls.

For a computer system, the human user or system administrator is asked to take up the accountability function. He or she should evaluate the log files. Does it work as expected? We can also imagine a situation in which this function is taken up by a software module that automatically evaluates the audit trail at frequent intervals; see, e.g., Continuous Control Monitoring (Alles et al. 2006; Kuhn and Sutton 2010). It remains an open question whether we can really speak about accountability or "assurance" in this case or whether it is only repeated verification dressed up in an accountability metaphor.

The principle of accountability states that an agent can be held accountable for the consequences of his or her actions or projects (Turilli and Floridi 2009). In that sense, accountability is similar to the moral notion of responsibility and the legal notion of liability. Based on the philosophical and legal literature (Duff 2007; Hart 1968), it is possible to identify a number of necessary conditions for attributing accountability. Accountability can be set apart from blame and from liability, which are all related to the umbrella notion of responsibility. Blameworthiness is the stronger notion. Accountability involves the obligation to justify one's actions. But such answerability does not necessarily imply blame. Ignorance or coercion can be valid excuses. Liability narrows the notion to a legal perspective. One can be blamed but not liable, for instance, when a contract explicitly excludes liability. Conversely, one can be liable for damages but not blamed, for instance, when everything was done to prevent disaster.

To summarize, according to Van De Poel (2011), agents are only accountable for their actions in case each of the following conditions are met:

(a) Capacity. The agent must be able to act responsibly. This includes the difficult issue of "free will." Conversely, when the agent is under pressure or coerced into doing something, or when he or she is mentally or physically disabled to act as expected, he or she is no longer held accountable.
(b) Wrongdoing. Something must be going wrong. This means that the agent fails to live up to its responsibilities to avoid some undesired condition *X*. Or, under a different account, it means that the agent transgressed some duty *D*.
(c) Causality. The agent is instrumental in causing some undesired condition *X*. Either this means that the agent is generally able to influence occurrence of *X* or that some transgression of duty *D* causes *X* to occur.

Accountability is essentially a property of people or organizations. Does it make sense to say that one can "design for accountability?" And what type of artifact is being designed in that case? For the notion to make sense, we must assume that the behavior of people in organizations can – to some extend – be designed. This is the topic of management control (Merchant 1998): how to design an organization and its procedures in such a way that its employees' behavior can be "controlled?"

Nowadays, much of the organizational roles, workflows, and procedures that restrict people's behavior in organizations are embedded and facilitated by information systems. They are designed. But also the control measures themselves are being designed, as are the ways of recording and processing information about behavior. These "measures of internal control" are the topic of this chapter. The definition runs as follows: "Internal control is a process, effected by an entity's board of directors, management and other personnel, designed to provide reasonable assurance regarding the achievement of objectives in the following categories: (i) Effectiveness and efficiency of operations, (ii) Reliability of financial reporting, and (iii) Compliance with applicable laws and regulations" (Coso 1992, p. 13). The notion of internal control originates in financial accounting, but based on the original COSO report and the later COSO ERM framework for risk management (Coso 2004), it has become widespread. Power (1997, 2007) has argued convincingly that this accounting perspective often produces a rather mechanistic notion of risk avoidance and control. For such mechanisms, it certainly makes sense to ask to what extent the design incorporates the values of accountability and transparency. After all, sometimes it does go wrong. In many of the accounting scandals of the 1990s, such as Enron or Parmalat, or in the demise of Lehman brothers in 2008, people were found to hide behind a formalistic approach to guidelines and procedures, instead of taking responsibility and assessing the real risks and reporting on them (Satava et al. 2006). One could argue that in such cases, the governance structure of the risk management procedures was not adequate. It was badly designed, because it only created traceability rather than facilitating a critical dialogue.

To summarize, we can say that it does make sense to "design for accountability." What is being designed, in that case, is the system of internal control measures, as well as business processes and information systems in which these controls are embedded. Obviously, other aspects of organizations, such as corporate culture (Hofstede, et al. 1990) or more specifically the risk culture (Power et al. 2013), also affect behavior. But we cannot properly say that these "soft" aspects of the control environment are designed. Instead they develop over time. They can at best be altered or facilitated.

Who is being held accountable? If we look at individual employees, designing for accountability would mean that conditions (a)–(c) hold for individuals. That would mean that employees have real responsibilities and do not just act to follow procedures. It also means that responsibilities and duties of different people are clearly delineated and that it is known who did what and what the consequences are of these actions. This last aspect involves an audit trail: all effects of actions must be traceable to an individual person. But it is crucial that at some point someone uses the audit trail to evaluate the behavior and confront employees with the consequences of their actions, be it good or bad.

If we look instead at accountability of the entire organization, we get a different picture. First, it means that evidence is collected to justify decisions: minutes of board meetings, requests and orders, or log files of transactions. Also the conditions under which the evidence is generated, processed, and stored matter. Information integrity must be assured, before it can function as reliable evidence. Integrity of information involves the properties of correctness (records correspond to reality), completeness (all relevant aspects of reality are recorded), timeliness (available in time for the purpose at hand), and validity (processed according to policies and procedures) (Boritz 2005). Important conditions to ensure integrity of information are good IT management, segregation of duties, access control, and basic security. Second, it means that some internal or external opposition is created, for instance, from a separate risk function or internal audit function reporting directly to the board of directors or from the external auditor or regulator. They must use this evidence to hold the management accountable. Typically, this is the hardest part (Power et al. 2013).

## Transparency

For transparency, a similar discussion can be started. If the notion applies to organizations, what does it mean to say that one is "designing for transparency?" And if it also applies to systems, what kinds of systems are being designed, when one is designing for transparency?

A transparent organization is one that has a tendency to be open in communication, unless there are good reasons not to. Transparency is the opposite of secrecy. Generally, transparency is believed to be a good thing. This is apparent, for example, in the movement for "open data," where government agencies are urged to "open up" the data they control (Zuiderwijk and Janssen 2014).

What is being designed? In the case of open data, it is about the policies and guidelines according to which officials can decide whether some specific database may be "opened up" for the public. A special example of such a policy is Executive Order 13526 about Classified National Security Information, which President Obama signed in the beginning of his first term. The purpose of this order was to declassify "old" government secrets and make them available to researchers and historians. In practice, such declassification programs face severe difficulties (Aftergood 2009). Government agencies like the CIA are often reluctant to declassify, for several reasons. In addition to genuine national security interests, Aftergood talks about bureaucratic secrecy. This is the tendency of bureaucracies to collect secrets, more than strictly needed. Apparently a bureaucracy creates incentives for officials to collect evidence but never to discharge. A third reason is "political secrecy," the tendency to use classification power for political purposes. "It exploits the generally accepted legitimacy of genuine national security interests in order to advance a self-serving agenda, to evade controversy, or to thwart accountability" (Aftergood 2009, p. 403).

Transparency as a property of organizations is impossible to design. As we have seen, transparency appears to be more about corporate culture and accepted practices than about formal procedures. Nevertheless, transparency can be facilitated, for instance, by providing infrastructures that make it easier to share and evaluate data.

What about transparency as a system property? Transparency does occur as a property of information systems, in several forms. Making explicit and comprehensible which rules are implemented in the system and how the information is produced is called procedural transparency (Weber 2008). An example of procedural transparency can be found in the edit history of Wikipedia and other Wikis. To increase reliability of information, anyone can examine the edit history and see who has edited, deleted, or added information (Suh et al. 2008). This should increase trust in the reliability of a source.

In electronic government research, transparency is also concerned with the (secured) disclosure of information to empower citizens to make better informed choices (Leite and Cappelli 2010). Information transparency can then be defined as "the degree to which information is available to outsiders that enables them to have informed voice in decisions and/or to assess the decisions made by insiders" (Florini 2007, p. 5).

In the development of computational models or computer simulations, transparency is concerned with making explicit the assumptions or values that are built into the model. To support users of the models to make informed decisions and develop an appropriate level of confidence in its output, it is important to explain their working (Fleischmann and Wallace 2009; Friedman et al. 2006). In that case, transparency is defined as the capacity of a model to be understood by the users of the model (Fleischmann and Wallace 2009). To ensure transparency, the model's assumptions about reality and values should be made explicit and testable.

Leite and Capelli (2010) discuss work on software transparency: all functions are disclosed to users. In human-computer interaction, an interface design is called

transparent, when it supports users in forming a correct mental representation of the working of the system. For example, the metaphor of a trash can on a desktop suggest that deleted files can be retrieved, until the trash is emptied. According to Norman (1998), a good interface should make the supporting computer technology disappear; what remains is mere functionality. This should make it easier to understand, learn, and use computer applications.

Transparency is also mentioned in the debate on open-source code: making the code available is the ultimate form of transparency. The underlying motivation is that scrutiny by the public will help to detect possible flaws and eventually increase quality of software. The same holds true for secure systems engineering. There is no such thing as "security by obscurity." Rather, security should be provided by the algorithms or the strength of the keys themselves (Schneier 2000). By opening up the source code of the algorithms, they can be tested and improved, if necessary.

Summarizing, depending on the context and use of the system, various kinds of transparency have been addressed in information systems research. From the perspective of those who gain access to information, transparency depends on factors such as the availability of information, its comprehensibility, accessibility, and how it supports the user's decision-making process (Turilli and Floridi 2009). When related to accountability, both transparency of information and procedural transparency are relevant. To enable a principal to trust and use information that is produced by a system on behalf of the agent, the way the information is generated must be visible: a black box would not suffice.

In general, information and communication technology is argued to facilitate accountability and transparency (Bannister and Connolly 2011). However, when accountability and transparency are improperly implemented, results can be disappointing and even contradictory. In the next section, we describe some undesirable consequences of designing for accountability and transparency.

## Controversy and Threats

Accountability and transparency are facilitated by ICT. ICT makes it easier to share and spread information within corporations and among their stakeholders. Vaccaro and Madsen (2009) argue that the use of ICT has realized social modifications and transformations in stakeholder relationships. Internet-based technologies such as e-mail, corporate websites, blogs, and online communities support intensive two-way information exchange between a firm and its stakeholders. Experiences acquired in this way can be used to modify the business practice and to make organizations more transparent, accountable, and socially responsible. Especially in business research, putting more information in the hands of stakeholders is assumed to force corporations to become more honest, fair, and accountable. But ICT can do little for benevolence and nothing for openness or empowerment if those in power do not want these things to happen (Bannister and Connolly 2011). Or as Elia (2009, p. 147) puts it, "technology may enable transparency but technology cannot guide it." Transparency through ICT increases trust because it leaves less to trust

(Bannister and Connolly 2011). However, in the context of e-government, trust in public processes cannot be delivered by technology alone; the basic structures of government need to be changed too (Bannister and Connolly 2011). Transparency stems from social interaction, and it is therefore a political, not a technical, issue (Menéndez-Viso 2009).

There are all kinds of political issues associated with transparency and accountability. For example, in the case of corporate governance, information may be "strategically disclosed." Corporations are not always willing to disclose information that will hold them accountable; often positive aspects are emphasized (Hess 2007). Elia subscribes these findings and argues that "information technology rarely creates access to information that a company has not vetted for public consumption" (Elia 2009, p. 148). Individuals also have strong motives against strategically disclosing information about their behavior (Florini 2007). One is that secrecy provides some insulation against being accused of making a mistake. It is much easier for an official to parry criticism when incriminating information remains secret. A second incentive is that secrecy provides the opportunity to uphold relationships with special interests. It is more difficult to maintain profitable relationships when financial transactions and the decision-making process are transparent.

There are circumstances under which secrecy is understandable and defendable. Bok (1983) lists the protection of personal identity, protection of plans before they have had the chance to be executed, and protection of property as legitimate reasons for justifying secrecy or, at least, for the right to control the choice between openness and secrecy. But even in a democracy, the right to make and keep things hidden is often misused to cover up political power struggles or bureaucratic failures, as we have seen in the example of Obama's declassification act (Aftergood 2009). In general, transparency is therefore considered to be preferred.

So if warranted transparency is a good thing, how should it be exercised? Making all data accessible at once is not very effective. Consider the Wikileaks cases, which required journalists to interpret them before they became "news." In fact, an excess of information makes accountability harder. It limits transparency, as it may be easy to drown someone in reports and hide behind an overwhelming amount of data presented through dazzling technological means (Menéndez-Viso 2009). Accountability requires useful and relevant information, not more information. To provide relevant information, it is particularly important to understand the information needs of the systems' users (Bannister and Connolly 2011). For example, shareholders want to be confident about the quality of the systems, processes, and competencies that deliver the information in the accountability report and underpin the organization's performance and commitments (Dando and Swift 2003).

Besides selective disclosure or over-disclosure, also false information may be distributed. The anonymity of the Internet is exploited to intentionally or unintentionally post erroneous information to manipulate the public opinion (Vaccaro and Madsen 2009). Consider, for example, restaurant reviews, which can be artificially "lifted." Even credible sources and full transparency may produce

incomplete, misleading, or untrue information. Therefore, it is essential to also acknowledge expertise, verifiability, and credibility (Santana and Wood 2009). But even when there is a genuine intention to achieve accountability, full and transparent disclosure of information may not be achieved because of the costs involved. Collecting, organizing, and disseminating information requires time, effort, and money. Agents will reveal information up to the point where the benefit from disclosure equals the costs. Typically this point is reached before full disclosure (Vishwanath and Kaufmann 2001).

To summarize, transparency can be too much of a good thing. Transparency mediated by information and communication technology runs a risk of being shallow, arbitrary, and biased toward corporate interests (Elia 2009). In this way transparency only creates apparent accountability. Partial or superficial transparency is considered to be more damaging than none at all (Elia 2009; Hess 2007). Providing limited transparency may be a technique to avoid disclosures which are truly relevant to stakeholder interests (Elia 2009). They may divert attention away from more effective means of accountability (Hess 2007).

## Designing for Accountability and Transparency

What does it mean to design for accountability? Or what does it mean to design for transparency? In the previous section, we discussed these questions in general. We will now make it specific in order to demonstrate – by example – that answering these questions makes sense. In addition we will explain some of the design issues to be addressed.

What kinds of systems can be designed for accountability or transparency? Here we will assume that we are designing and developing an information system, in its broadest (socio-technical) sense: a collection of devices, people, and procedures for collecting, storing, processing, retrieving, and disseminating information, for a particular purpose. In case of accountability, we are designing essentially the collection and storage of information about the actions of the agent, which may then be used as evidence in an accountability relation, i.e., for the purpose of providing justification to the principal. In case of transparency, we are designing essentially a dissemination policy. We assume the evidence is there, but unless there are mechanisms to facilitate dissemination, it will be kept private. As we have seen, transparency may be restricted by legitimate secrets, but this requires an explicit classification decision.

Consider an agent, which is accountable to a forum. As announced in the introduction, we focus here on compliance reporting. The agent is a company, and the forum consists of a regulator on behalf of the general public. We will now present a simplified architecture of a reporting system that could facilitate such an accountability relation. What we are monitoring are the primary processes of the agent but also whether the agent is “in control”: is the agent achieving its control objectives and if not, is it adjusting the process? Therefore, the internal control processes need to be monitored too. All of this generates evidence, which is stored,

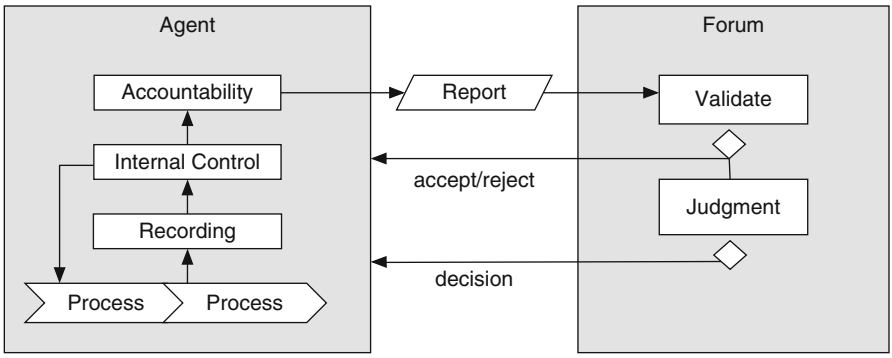



**Fig. 2** Simplified compliance reporting architecture

but also compiled into an accountability report and sent to the forum. The forum will then evaluate the report and pass judgment (Fig. 2).

Based on this reporting architecture, we can list a number of requirements. First we list requirements that are related to accountability. Next we will list some requirements specifically about transparency:

(A1) Recording evidence. First we have to make sure there is reliable data at all. This concerns the initial point of recording of behavior (Blokdijk et al. 1995). Data should be recorded from an independent source or if that is impossible, automatically. For raw data to become evidence, it must be interpreted as representing specific events or propositions, which are meaningful in the legal context. It should be evidence of something. Moreover, the evidence should be recorded in a read-only storage device that allows for easy retrieval. The read-only device should make sure the evidence is not lost or manipulated.

(A2) Segregation of duties. The objectivity of the evidence is based on the principle of segregation of duties. According to this age-old accounting principle, the organizational roles of authorizing a decision, executing it, and recording its effects should be executed by independent actors (Romney and Steinbart 2006; Starreveld et al. 1994). Even stronger is the principle of countervailing interests. Parties to a commercial transaction generally have countervailing interests. Therefore, they will scrutinize the accuracy and completeness of data in commercial trade documents. For instance, buyer and seller will cross verify the price in an invoice. In government, countervailing interests are artificially created, for instance, between front office and back office.

(A3) How much data should be recorded? Generally, what should be recorded is an audit trail: a trace of decisions, which allows others to recreate the behavior after the fact. If too much is recorded, auditors will drown in details and relevant material will not be found. Moreover, there may be performance limitations. If too little is recorded, the data is useless as evidence.

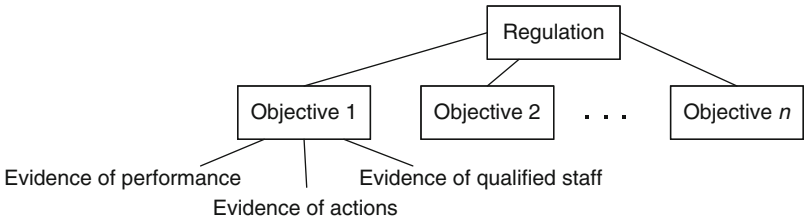



**Fig. 3** Determining kinds of evidence

The difficulty is that future information needs are unknown. The trade-off can be solved by a risk analysis, which should involve all parties interested in accountability: risk officer, compliance officer, business representative, and system administrator. They should evaluate the risk of missing evidence against performance losses.

(A4) What should be recorded? Generally we are interested in the effectiveness of measures, in achieving meeting specific control objectives, often based on a more general regulation or guideline. There are several ways of defining controls, based on different performance indicators (Eisenhardt 1985). We can record evidence of the actual performance, if that can be measured at all, we can record evidence of the actions that were undertaken, or we can record evidence of the necessary side conditions, such as qualified staff. Consider, for example, the monitoring of hygiene standards in the food industry. Performance can be measured directly by taking samples of the E. coli. The amount of bacteria per sample is considered a good indicator of relative hygiene. We can also measure the cleaning actions, whether they are indeed executed according to schedule. Finally, we can verify if all staff has followed the hygiene-at-work course (Fig. 3).

(A5) Being in control. Instead of collecting evidence of the primary processes, in many cases it makes sense to look at the maturity level of the internal controls. Here we can test whether the company is actually "in control." For example, we can verify that whenever there are incidents, there is a follow-up, and the effectiveness of the follow-up is tested. This kind of thinking is very common in quality management and risk management (Coso 2004). A typical representative of this approach is the plan-do check-act cycle (Deming 1986).

(A6) Critical Forum. As we argued above, evidence of past actions is not enough to bring about accountability. Someone must critically examine the evidence and confront those who are responsible with the consequences of their actions. This is the role of the forum (Bovens 2007). Organizing a critical opposition is key to accountability. This aspect is part of the governance structure of an organization. For example, in risk management, a separate risk function is created, which reports directly to the board of directors to ensure independence of the business, whom it needs to challenge (Iia 2013).

The following requirements have to do with transparency, namely, with policies determining the dissemination of accountability evidence:

(T1) Reliable Reporting. An accountability report needs to be accurate and complete. That means it must be compiled on the basis of all the evidence generated in the processes discussed above and only on the basis of such evidence. As they perform a communicative function for some specific audience, reports should be coherent, informative, true, relevant, and not over-informative with respect to the information needs of the audience; compare the maxims of Grice (1975).

(T2) Audience. All reports are written for a specific audience, in this case the forum. Generally reports are compiled according to standards, which should be based on the information needs of the forum. Such information needs are again based on the decision criteria for the judgment that depends on the report. In practice this means that the one-size-fits-all approach to annual financial statements, which are supposed to serve the general public, is inappropriate. Different stakeholders have different information needs. Modern technology like XBRL makes it relatively easy to differentiate and generate different reports on the basis of the same evidence (Debreceny et al. 2009).

(T3) Statement. Accountability reports are more than a collection of evidence. They are compiled and accompanied by a statement of the management that the report is considered an accurate and complete representation of reality. This statement generates a commitment to the truth of the information; it also generates a sense of responsibility and in many cases also legal liability. For example, under the Sarbanes-Oxley Act, managers are personally liable for misstatements. This is supposed to have a deterrence effect.

The role of the external auditor is interesting in this respect. Financial statements are directed to the general public, including shareholders. The accountant does not play the role of "significant other" but merely provides assurance. The accountant performs an investigation of the records on which the statements are based, in order to attest to the accuracy and completeness of the financial statements, which are compiled by the company itself (Knechel, et al. 2007). In the architecture above, that means that the process "accountability" has been partly delegated to an independent agency.

Another interesting remark concerns the nature of evidence. Typically, depending on the context, some records or logs are said to "count as" evidence in a legal or institutional sense. This counts-as relation has been studied by John Searle (1995). Using such constitutive rules, we are creating our complex social environment, in which many facts are not brute facts but rather institutional facts. Generally, counts-as rules have the following form: in institutional context $S$, act or event $X$ counts as institutional fact or event $Y$. Consider, for example, a receipt. When I buy something, the paper receipt I am given counts as legal evidence that I have indeed bought and received the product for that price. Searle's theory shows that

evidence is essentially a social construction. It only makes sense within a specific institutional context, in our case defined by the forum. For example, the auditing term "test of operating effectiveness" has a different and much more limited meaning than the general meaning of the term "effectiveness." Operating effectiveness means that a control has been operational for the full duration of the period under investigation. It is not concerned with the effectiveness of the measure, whether it was successful in achieving its goals. Such conceptual differences between auditor and forum may hamper transparency.

## Value-Based Argumentation

As we have seen in the previous two sections, designing is essentially a matter of solving trade-offs. Consider issues such as: how much evidence? What is the cost of compliance? What is legitimate secrecy? We believe that such trade-offs can only be solved in a dialogue between stakeholders. Generally, different stakeholders will represent different interests. So it is important to have all parties present during such design workshops. There are many techniques to facilitate decision making in groups. Consider for example, the Delphi method, which tries to gather and evaluate innovative ideas from a group of experts. However, here we are dealing with a specific decision task: designing a system. Not just opinions matter but also the technical feasibility of the resulting system. Generally, the dialogue will be a mixture of technical arguments about facts, effectiveness of measures or feasibility, and more motivational arguments about the relative desirability of specific objectives and underlying social values.

Argumentation is an interactive process in which agents make assertions (claims) about a certain topic, which support or attack a certain conclusion or support or attack the assertions of the opponent. There is no single truth; what matters is the justification of the conclusion (Walton 1996). In previous research, we have developed *value-based argumentation theory*, a dialogue technique to make the values underlying decisions about system requirements explicit (Burgemeestre et al. 2011, 2013). The approach is based on an earlier approach to value-based argumentation used for practical reasoning (Atkinson and Bench-Capon 2007). We believe that the structured nature of the argumentation framework with its claims and counterattacks closely resembles an audit process as we encounter it in practice. In the remainder, we will give a brief exposition of the approach, enough to show that there are indeed techniques that can help facilitate design processes, which take values like accountability and transparency into account.

Walton (1996) organizes practical reasoning in terms of argument schemes and critical questions. An argument scheme presents an initial assertion in favor of its conclusion. Now it is up to the opponent to try and disprove the assertion. The opponent can challenge the claims of the proponent by asking so-called critical questions. Originally, the argumentation scheme uses means-end reasoning: what kinds of actions should we perform in order to reach our goals? This already captures debates about effectiveness and about alternatives. But how are goals

justified? The answer is by social values. Perelman (1980) indicates that social values can account for the fact that people may disagree upon an issue even though it would seem to be rational. In the business world, consider values like profit, safety, or quality. Such values are embedded in the corporate culture (Hofstede et al. 1990). For example, a culture which values short-term profits over security – as apparent from payment and incentive schemes – will be more likely to lead to behavior which violates a security norm than a culture which values security over profits.

Atkinson et al. (2006) and Atkinson and Bench-Capon (2007) have adapted Walton's argument scheme and added social values. In our work on designing security systems, we have in turn adapted Atkinson et al.'s argument scheme to the requirements engineering domain (Burgemeestre et al. 2013). This has resulted in the following argumentation scheme:

(AS) In the current system $S_1$,
we should implement system component $C$,
resulting in a new system $S_2$ which meets requirements $R$,
which will realize objective $O$
and will promote value $V$.

An argument scheme like AS asserts an initial position. Opponents may then ask critical questions (CQ), trying to undermine the assumptions underlying the argumentation. Atkinson et al. (2007) provide an extensive list of critical questions, challenging the description $S_1$ of the facts of the case, the effectiveness of the choice of action $C$, or the legitimacy of the social value $V$. We have adapted this list of questions (Burgemeestre et al. 2013). In our exposition here we use a simplified version. Note that a further simplification would be possible if we would connect values to requirements directly, without the additional layer of objectives. However, in practice the additional layer is needed to deal with objectives, which are more general than system requirements but which are not values, for example, regulatory objectives.

CQ1. Is the current system $S_1$ well described?
CQ2. Will implementing component $C$ result in a system which meets requirements $R$?
CQ3. (a) Are requirements $R$ sufficient to achieve objective $O$?
(b) Are requirements $R$ necessary to achieve $O$?
CQ4. Does new system $S_2$ promote the value $V$?
CQ5. Are there alternative systems $S'$ that meet requirements $R$ and are there alternative sets of requirements $R'$ that achieve objective $O$ and promote value $V$?
CQ6. Does component $C$ have a negative side-effect $N$, which demotes value $V$ or demotes another value $W$?
CQ7. Is implementing component $C$, meeting $R$, and achieving $O$ feasible?
CQ8. Is value $V$ a justifiable value?

Critical question CQ1 is about the accuracy of the current system description. CQ2 is about effectiveness of the component; CQ3 is about effectiveness of the requirements specification, in meeting the objective. This involves two questions: adequacy refers to the question whether $R$ is enough to achieve $O$. Necessity refers to the question whether no requirements can be left out. This is related to efficiency, in a way. One wants to achieve objective $O$ with as little demands as possible. Note that there may be several ways of achieving the same objective. CQ4 considers whether the resulting system as a whole will promote the value. CQ5 considers alternative solutions. CQ6 considers possible negative side effects. Typically, the costs of an investment are listed here. CQ7 considers feasibility of the solution. In other words, are the assumptions on which the reasoning is based warranted? Finally, CQ8 considers the relative worth of the value itself.

## Dependency Graphs

In addition to this dialogue technique with argumentation schemes and critical questions, we have also developed a graphical notation (Burgemeestre et al. 2011). The diagrams are called dependency graphs. They are based on AND/OR graphs and depict the dependency relationships from general objectives, requirements, and other system properties to specific system components (Fabian et al. 2010). They are similar to diagrams in the i* modeling approach (Yu 1997). We can settle trade-offs between objectives, requirements, and properties by linking objectives to social values and thus determining their relative priority.

▶ **Definition 1** (Dependency Graph) Given a set of system components $C$, properties $P$ (including requirements $R$), objectives $O$, and values $V$, a dependency graph is defined as a directed acyclic graph $\langle N, A+, A- \rangle$, where nodes $n_1, n_2, \ldots$ can be any element of $C$, $P$, $O$, or $V$. There are two kinds of arcs. Positive arcs $n_1 \rightarrow n_2$ mean that $n_1$ *contributes to achieving* $n_2$. Negative arcs $n_1 \rightarrow^{-} n_2$ mean that $n_1$ *contributes to* not *achieving* $n_2$.

For example, the graph in Fig. 3 may depict the contents of a dialogue about security measures. It is agreed among the participants that a system with a *weak password* policy contributes to the value *usability*, whereas a *strong password* policy promotes *security*. A *strong password* policy has the property of being *hard to remember* for users. This property in turn negatively affects the value *usability*. Depending on the situation, different choices will be made. In a school, usability probably outweighs security, but not in a bank (Fig. 4).

By exchanging arguments, dialogue participants increase their shared knowledge. They learn things about the system and about each other. The dependency graphs can be understood as a means to capture the shared knowledge about the system at a certain point in the dialogue. For example, a designer may argue that the property of being *hard to remember* also negatively affects *security*, as it may

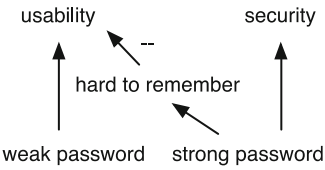



**Fig. 4** Simple example of a dependency graph

lead to people writing their passwords down. Such a dialogue move would add a second negative arrow: *hard to remember* $\rightarrow^{-}$ *security*.

## Alternatives

In the field of requirements engineering, there are other approaches that also take nonfunctional requirements such as accountability, security, or privacy into account. In particular, there is the concept of soft goals: "subjective and unstructured or ill-defined qualities" (Bresciani et al. 2004; Yu 1997). Just like values, soft goals can be used to prioritize, evaluate, and compare other more specific requirements. However, these approaches do not explicitly deal with the fact that there may be different stakeholders with possibly opposite objectives, which need to be decided upon. In other words, they do not take the dialogue aspect seriously. A recent exception is presented by Prakken et al. (2013), who use argumentation to facilitate risk analysis in designing secure systems.

In the field of AI and law, there is a whole body of literature on the use of formal argumentation, for example, to capture legal reasoning about evidence; see, e.g., (Bex et al. 2003). Value-based argumentation also derives from that tradition (Atkinson et al. 2006). Part of the research on argumentation has branched off and developed relatively abstract mathematical models, called Abstract Dialectical Frameworks (Brewka et al. 2013). Now it turns out that these ADFs are very similar to the dependency graphs presented here. An ADF is essentially a graph, where nodes represent statements and links represent either positive (+) or negative (−) contributions to reaching a specific conclusion.

## Experiences: Cooperation Between Regulators and Software Providers

Currently the quality of cash registers and point-of-sales (POS) systems that are available on the market is highly diverse. Enterprises can buy anything from second-hand cash registers to high-end POS systems that are integrated with back-office information systems. Especially for the lower market, segment competition is based on price rather than quality. Furthermore, cash register vendors feel pressured to satisfy customer requests to implement mechanisms that might enable fraud. A famous example is the testing mode, which allows managers to set the internal cash flow counter, at the heart of a point-of-sales system, back to zero.

**Table 1** Example of a control objective and corresponding norms

| Control objective 1: register all events | |
|---|---|
| Nr | Norm |
| 1 | All events occurring on the POS system during the formalization phase are being registered |
| 2 | From the start of the formalization phase, data about transactions are being stored |
| 3 | Corrections are processed without altering the original transaction. Additional corrections must be traceable to the original transaction with an audit trail |

By contrast, point-of-sales systems that do have a reliable counter can be used as an independent source of evidence about the incoming cash flow (recall A1). For this reason, the testing mode is nicknamed the "Marbella button," because it allows employees to divert revenue without being detected and book a nice holiday.

In this case study, we discuss the collaboration between the Dutch Tax and Customs Administration (DTCA) with vendors and developers of high-end POS systems to develop a quality mark. Such a mark will serve as a signal to differentiate high- from low-quality cash registers. In addition, a quality mark will increase awareness among businesses about the necessity for reliable point-of-sales systems. Note that corporate taxes are generally calculated as a percentage of the company's revenues. VAT is calculated as a percentage of the sales. This explains the involvement of the tax administration. In the words of director-general Peter Veld: "This will make it easier for businesses to choose a cash register that will not hide revenue from the tax office."[1] For the quality mark to become a success, norms or standards are needed to differentiate reliably between high- and low-quality cash systems. A quality mark should only be given to systems that do not enable fraud.

DTCA and market participants have jointly developed a set of norms for systems to qualify for the quality mark, called "Keurmerk betrouwbare afrekensystemen" (quality mark reliable point-of-sales systems). The objective is that "a reliable POS system should be able to provide a correct, complete and continuously reliable understanding of all the transactions and actions performed with and registered by the system." The norms are set up in a principle-based fashion, to make them applicable to a wide variety of POS systems on the market and to make them flexible enough to account for future developments. There are two versions, one for closed systems – these conform to all constraints of DTCA – and one for systems that can be closed, given the right settings. These systems are delivered with a delivery statement, which states that at the point of delivery, the system was configured in such a way that the system conformed to all constraints.

The norms are grouped into four main control objectives: (1) register all events, (2) preserve integrity of registrations, (3) secure storage of registrations, and (4) provide comprehensible and reliable reporting. Norms then prescribe specific requirements that a POS system must meet in order to fulfill the control objective. As an example, we will now discuss the first control objective in more detail (Table 1).

[1] http://www.keurmerkafrekensystemen.nl/, last accessed 6th of November 2014.

The norms use a conceptual model of a sales transaction, consisting of the following phases: selection of the goods, showing financial data (e.g., price), formalization of the transaction, confirmation, and payment. Formalization is the phase during which all events must be recorded. Formalization starts at the latest when data of a financial nature are being shown. The overarching control objective is that all events performed with a POS system are registered. To be able to determine whether transactions are paid and registered in an accurate, complete, and timely manner, not only the transactions but also special actions like discounts, returns, canceled transactions, withdrawals, training sessions, and corrections should be identified and registered. To assure completeness of the audit trail, corrections should not delete or alter the original transaction data but rather reverse them by an additional credit booking.

## Example Argumentation

A well-known fraud scenario is called "fake sales." The cashier first scans or enters the product into the POS system and the transaction data is displayed. The customer pays the money to the cashier, takes the product, and leaves the shop. Instead of confirming the sales transaction (formalization) and registering the transaction data in the database, the cashier then cancels the transaction and steals the money. The following dialogue shows two participants of the working group, discussing controls to address the "fake sales" scenario.

A1 Suppose a POS system is used to show the customer an example offer. The transaction is canceled. The product is not sold. Should this event be registered by the system (CQ3a)?

B1 Even though the product is not delivered, it is a transaction because financial data is shown to the customer and all financial transactions should be registered.

A2 Currently the removal of an offer is registered as a "no sale."[2] All other details of the transaction are removed from the database.

B2 The "no sale" option enables fraud. When a customer does not request a receipt, the entrepreneur can choose the "no sale" option instead of confirming the sales transaction. The transaction data and revenues are not registered in the POS system database, and the entrepreneur can secretly collect the money (CQ6).

A3 But an honest entrepreneur will not abuse the "no sale" option to commit fraud.

B3 Scanning a product without registering the transaction can also form a risk for the entrepreneur. Employees can also steal money when details on sales transactions are not registered by the POS system (CQ6).

[2]A "no sale" is an action on the POS system that has no financial consequences, for example, opening the cash register to change bills into coins.

B4 To prevent abuse of the "no sale" option, one should keep a record of why the "no sale" action was used and by whom (CQ5).
A4 But writing a motivation for each canceled transaction disrupts the sales process (CQ7).
B5 Then at least metadata details on the "no sale" action should be registered in the audit trail. For example, the employee that performed the action, the time and date of the transaction, and the duration of the transaction (CQ5).
A5 Ok.

## Example Dependency Diagram

In this chapter we define the values, objectives, system properties (including requirements), and system components used to capture the essence of the dialogue. We model the dependencies between these concepts in a dependency graph.

Values:

$v_1$: Completeness
$v_2$: Profitability
$v_3$: Accountability
$v_4$: Usability

Objectives:

$o_1$: Register all (trans)actions
$o_2$: Steal money
$o_3$: Efficient sales process
$o_4$: Detect fraud

Properties:

$p_1$: All events during the formalization phase of a transaction are registered.
$p_2$: All actions are registered.
$p_3$: All corrections are registered.
$p_4$: Corrections do not alter transactions.
$p_5$: Register metadata about "no sale."

Components:

$c_1$: "No sale"
$c_2$: Motivation
$c_3$: Metadata

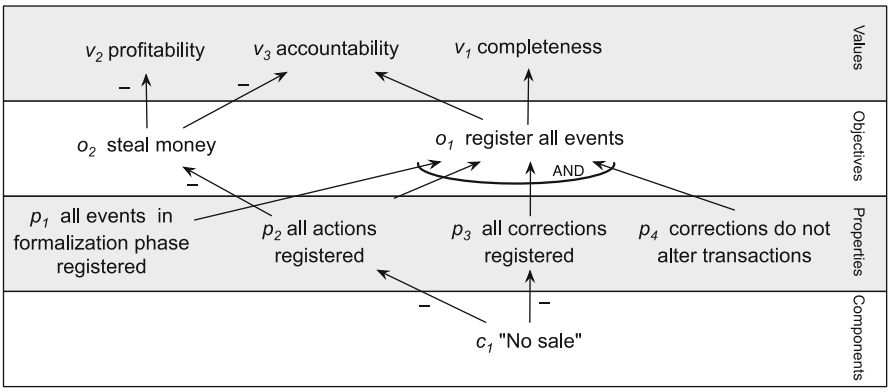



**Fig. 5** Dependency diagram "no sale"

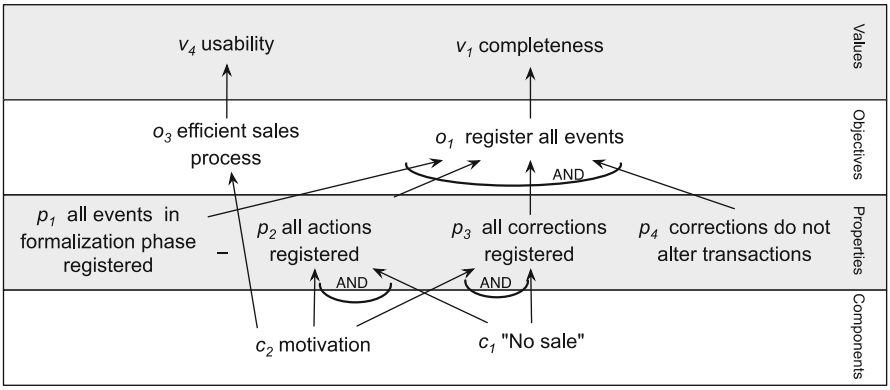



**Fig. 6** Dependency diagram "no sale" with a motivation

We will now visualize the dependencies between these concepts as expressed in the dialogue in a series of dependency diagrams (Fig. 5–7).

By using the "no sale" option, the transaction is not confirmed and the actions preceding the transaction are not registered. Therefore, objective $o_1$ of registering all events is not achieved and completeness is not promoted. Furthermore, the "no sale" option is used with the intention to steal money, which has a negative effect on profitability and accountability (Fig. 6).

When in addition some motivation must always be included that explains by whom and why the "no sale" option was used, we can say that all actions are in fact registered, $o_1$ is achieved, and completeness is promoted. However, having to include a motivation has a negative effect on the efficiency of the sales process, and therefore the usability of the system is demoted (Fig. 7).

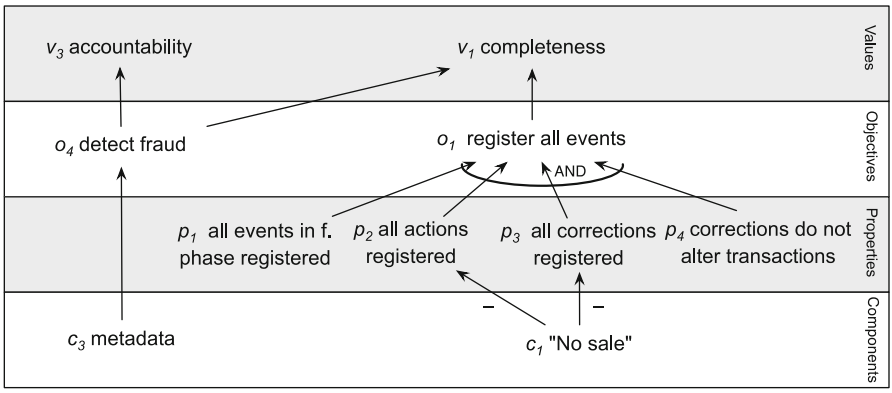



**Fig. 7** Dependency diagram "no sale" with metadata logging

In the end of the dialogue, registering metadata is proposed as an alternative compensating measure. In this situation, all actions are still not registered due to the "no sale" option, but instead metadata is registered that serves as an indicator to determine the completeness of transactions. By analyzing metadata, deviating patterns in the sales process can be found and fraud can often be detected. For example, some employees may have significantly more "no sales" activities. When such fraud can be detected, this feature has a positive influence on accountability and the completeness of transactions.

## Lessons Learned

The case study shows that we can use value-based argumentation to make trade-offs between different values, objectives, and system components explicit. From a dialogue we derive the participants' considerations on the design of reliable POS systems. By modeling the dependencies between these concepts, we can compare the advantages and disadvantages of a specific solution. As values are now explicitly linked to requirements and system components, we may conclude that our methodology does indeed provide a conceptual approach to take values into account in the design of information systems.

We also found that our approach has limitations, see also (Burgemeestre et al. 2011). The example of "fake sales" is rather simple. We can imagine that analyzing a more complex problem requires a lot more effort. The outcomes may be difficult to depict in dependency diagrams. Furthermore, human dialogues can be very unstructured, and therefore it might require considerable effort to capture them in terms of argumentation schemes and critical questions. Finally, as in all conceptual modeling, a lot depends on the modeling skills of the person making the diagrams. Little differences in the use of notation have a large effect.

## Conclusions

An accountability relationship exists when one agent performs actions and feels obliged to provide evidence of his or her actions to the forum: some significant other, such as a regulator, the general public, or the principal on whose behalf the actions are performed. In this chapter we focus in particular on the application of such a setting to compliance monitoring. In this case, the agent is a company; the forum is a regulator. The accountability reports are required as part of a self-regulation or co-regulation scheme. Systems designed for accountability make sure that relevant information about the agent's actions are recorded reliably and stored in a secure way, in such a way that all actions can be traced later. Provided that necessary precautions and guarantees are built into the way the information is being recorded and processed – internal controls – such information may then serve as evidence to justify decisions later.

Designing for accountability is therefore closely related to the question of what counts as evidence. But designing for accountability is more than implementing traceability. It also involves the organization of an effective opposition, either internal (independent risk function) or external (regulator; shareholders). This opposition must actively take on the role of a forum, use the audit trail to evaluate behavior, and confront agents with the consequences of their actions, either good or bad.

Designing for transparency amounts to designing a dissemination policy for the evidence, once it has been collected. Transparency can be restricted by illegitimate attempts to manipulate the dissemination process, for example, by causing information overload or by providing false information. Transparency can also be restricted by legitimate secrets. What is considered legitimate in a specific domain needs to be decided up front. Therefore, classification policies should be maintained that have disclosure as the default. For example, consider a system that would "automatically" reveal secrets after their expiry date.

As all design efforts, designing for accountability or transparency involves trade-offs. Social values like accountability or transparency may conflict with profitability, with safety and security or with mere tradition. In this chapter we have shown a dialogue technique for making such trade-offs explicit: value-based argumentation theory. The idea is that one stakeholder makes a proposal for a specific design, motivated by goals (design objectives) and underlying social values. The other stakeholders may then ask questions, challenge the underlying assumptions, or propose alternatives. The proposer is forced to defend the proposal, using arguments. Eventually, when the space of possible arguments is exhausted, the remaining proposal is best suited because it has survived critical scrutiny. The knowledge exchanged in the dialogue can be depicted in a dependency graph.

We have described a case study of the development of quality standards for point-of-sales systems, supported by the tax office. In the case study, we have demonstrated that designing for accountability is essentially a group effort, involving various stakeholders in opposite roles. Stakeholders challenge assumptions and provide clarification, thus increasing shared knowledge about the design and

improving the motivation for specific design choices. We have also shown that value-based argumentation is a technique that can in fact be used to analyze the trade-offs in the case and provide fruitful insights. In addition, we have used dependency graphs to provide graphical summaries of the dialogue outcomes.

Although we derive these conclusions on the basis of a relatively specific example, we believe that they are generic. Values are intimately connected to the identity of people. Also other engineering disciplines that are struggling with dilemmas concerning safety, sustainability, or health could therefore benefit from such critical dialogues about values. Once again, this requires a critical forum that is not afraid to challenge assumptions and a transparent attitude that prefers making the motivation for design choices explicit. Under such conditions, the designers can be held accountable for their design.

## References

Aftergood S (2009) Reducing government secrecy: finding what works. Yale Law Policy Rev 27:399–416

Alles M, Brennan G, Kogan A, Vasarhelyi M (2006) Continuous monitoring of business process controls: a pilot implementation of a continuous auditing system at Siemens. Int J Acc Inf Syst 7:137–161

American Accounting Association, C. o. B. A. C. C (1972) Report of the committee on basic auditing concepts. Acc Rev XLVII:14–74

Atkinson K, Bench-Capon T (2007) Practical reasoning as presumptive argumentation using action based alternating transition systems. Artif Intell 171:855–874

Atkinson K, Bench-Capon T, McBurney P (2006) Computational representation of practical argument. Synthese 152(2):157–206

Ayres I, Braithwaite J (1992) Responsive regulation: transcending the deregulation debate. Oxford University Press, New York

Bannister F, Connolly R (2011) Trust and transformational government: a proposed framework for research. Gov Inf Q 28(2):137–147. doi:10.1016/j.giq.2010.06.010

Bex FJ, Prakken H, Reed C, Walton DN (2003) Towards a formal account of reasoning about evidence: argumentation schemes and generalisations. Artif Intell Law 11:125–165

Black J (2002) Regulatory conversations. J Law Soc 29(1):163–196

Blokdijk JH, Drieënhuizen F, Wallage PH (1995) Reflections on auditing theory, a contribution from the Netherlands. Limperg Instituut, Amsterdam

Bok S (1983) Secrets: on the ethics of concealment and revelation. Pantheon Books, New York

Boritz JE (2005) IS practitioners' views on core concepts of information integrity. Int J Acc Inf Syst 6(4):260–279

Bovens M (2007) Analysing and assessing accountability: a conceptual framework. Eu Law J 13(4):447–468

Bovens M et al (2005) Public accountability. In: Ferlie E (ed) The Oxford handbook of public management. Oxford University Press, Oxford, UK

Breaux T, Anton A (2008) Analyzing regulatory rules for privacy and security requirements. IEEE Trans Softw Eng 34(1):5–20. doi:10.1109/tse.2007.70746

Bresciani P, Perini A, Giorgini P, Giunchiglia F, Mylopoulos J (2004) Tropos: an agent-oriented software development methodology. J Auto Agent Multi-Agent Sys 8:203–236

Breu R, Hafner M, Innerhofer-Oberperfler F, Wozak F (2008) Model-driven security engineering of service oriented systems. Paper presented at the information systems and e-Business Technologies (UNISCON'08)

Brewka G, Strass H, Ellmauthaler S, Wallner JP, Woltran S (2013) Abstract dialectical frameworks revisited. Paper presented at the 23rd international joint conference on artificial intelligence (IJCAI 2013), Being

Burgemeestre B, Hulstijn J, Tan Y-H (2011) Value-based argumentation for justifying compliance. Artif Intell Law 19(2–3):149–186

Burgemeestre B, Hulstijn J, Tan Y-H (2013) Value-based argumentation for designing and auditing security measures. Ethics Inf Technol 15:153–171

Chopra AK, Singh MP (2014) The thing itself speaks: accountability as a foundation for requirements in sociotechnical systems. In: Amyot D, Antón AI, Breaux TD, Massey AK, Siena A (eds) IEEE 7th international workshop on requirements engineering and law (RELAW 2014), Karlskrona, pp 22–22

COSO (1992) Internal control – integrated framework. Committee of Sponsoring Organizations of the Treadway Commission, New York

COSO (2004) Enterprise risk management – integrated framework. Committee of Sponsoring Organizations of the Treadway Commission, New York

Dando N, Swift T (2003) Transparency and assurance minding the credibility gap. J Bus Ethics 44(2):195–200. doi:10.1023/a:1023351816790

Day P, Klein R (1987) Accountabilities: five public services. Tavistock, London

Debreceny R, Felden C, Ochocki B, Piechocki M (2009) XBRL for interactive data: engineering the information value chain. Springer, Berlin

Deming WE (1986) Out of the crisis. MIT Center for Advanced Engineering Study, Cambridge

Dubnick MJ (2003) Accountability and ethics: reconsidering the relationships. Int J Org Theory Behav 6:405–441

Dubois E, Mouratidis H (2010) Guest editorial: security requirements engineering: past, present and future. Requir Eng 15:1–5

Duff A (2007) Answering for crime: responsibility and liability in the criminal law. Hart Publishing, Oxford

Eisenhardt KM (1985) Control: organizational and economic approaches. Manag Sci 31(2):134–149

Eisenhardt KM (1989) Agency theory: an assessment and review. Acad Manage Rev 14(1):57–74

Elia J (2009) Transparency rights, technology, and trust. Ethics Inf Technol 11(2):145–153. doi:10.1007/s10676-009-9192-z

Eriksén S (2002) Designing for accountability. Paper presented at the NordiCHI 2002, second Nordic conference on human-computer interaction, tradition and transcendence, Århus

Fabian B, Gürses S, Heisel M, Santen T, Schmidt H (2010) A comparison of security requirements engineering methods. Requir Eng 15:7–40

Fleischmann KR, Wallace WA (2009) Ensuring transparency in computational modeling. Commun ACM 52(3):131–134. doi:10.1145/1467247.1467278

Flint D (1988) Philosophy and principles of auditing: an introduction. Macmillan, London

Florini A (2007) Introduction. The battle over transparency. In: Florini A (ed) The right to know. Transparency for an open world. Columbia University Press, New York

Friedman B, Peter H, Kahn J (1992) Human agency and responsible computing: implications for computer system design. J Syst Softw 17(1):7–14. doi:10.1016/0164-1212(92)90075-u

Friedman B, Kahn PH Jr, Borning A (2006) Value sensitive design and information systems. In: Zhang P, Galletta D (eds) Human-computer interaction in management information systems: applications, vol 6. M.E. Sharpe, New York, pp 348–372

Grice HP (1975) Logic and conversation. Syntax Semant 3

Hart HLA (1968) Punishment and responsibility: essays in the philosophy of law. Clarendon, Oxford

Heckman C, Roetter A (1999) Designing government agents for constitutional compliance. Paper presented at the proceedings of the third annual conference on autonomous agents, Seattle

Hess D (2007) Social reporting and new governance regulation: the prospects of achieving corporate accountability through transparency. Bus Ethics Q 17(3):453–476

Hofstede G, Neuijen B, Ohayv DD, Sanders G (1990) Measuring organizational cultures: a qualitative and quantitative study. Adm Sci Q 35(2):286–316
IIA (2013) The three lines of defense in effective risk management and control. IIA Position Papers, The Institute of Internal Auditors (IIA)
Johnson DG, Mulvey JM (1995) Accountability and computer decision systems. Commun ACM 38(12):58–64. doi:10.1145/219663.219682
Knechel W, Salterio S, Ballou B (2007) Auditing: assurance and risk, 3rd edn. Thomson Learning, Cincinatti
Korobkin RB (2000) Behavioral analysis and legal form: rules vs. principles revisited. Oregon Law Rev 79(1):23–60
Kuhn JR, Sutton SG (2010) Continuous auditing in ERP system environments: the current state and future directions. J Inf Syst 24(1):91–112
Leite JC, Cappelli C (2010) Software transparency. Bus Inf Syst Eng 2(3):127–139. doi:10.1007/s12599-010-0102-z
Menéndez-Viso A (2009) Black and white transparency: contradictions of a moral metaphor. Ethics Inf Technol 11(2):155–162. doi:10.1007/s10676-009-9194-x
Merchant KA (1998) Modern management control systems, text & cases. Prentice Hall, Upper Saddle River
Nissenbaum H (1994) Computing and accountability. Commun ACM 37(1):72–80. doi:10.1145/175222.175228
Norman DA (1998) The invisible computer. MIT Press, Cambridge
Perelman C (1980) Justice law and argument. D. Reidel Publishing, Dordrecht
Power M (1997) The audit society: rituals of verification. Oxford University Press, Oxford
Power M (2007) Organized uncertainty: designing a world of risk management. Oxford University Press, Oxford
Power M (2009) The risk management of nothing. Acc Organ Soc 34:849–855
Power M, Ashby S, Palermo T (2013) Risk culture in financial organisations. London School of Economics, London
Prakken H, Ionita D, Wieringa R (2013) Risk assessment as an argumentation game. Paper presented at the 14th international workshop on computational logic in multi-agent systems (CLIMA XIV)
Rees J (1988) Self Regulation: an effective alternative to direct regulation by OSHA? Pol Stud J 16(3):602–614
Romney MB, Steinbart PJ (2006) Accounting information systems, 10th edn. Prentice Hall, Upper Saddle River
Santana A, Wood D (2009) Transparency and social responsibility issues for Wikipedia. Ethics Inf Technol 11(2):133–144. doi:10.1007/s10676-009-9193-y
Satava D, Caldwell C, Richards L (2006) Ethics and the auditing culture: rethinking the foundation of accounting and auditing. J Bus Ethics 64:271–284
Schneier B (2000) Secrets and lies: digital security in a networked world. Wiley, New York
Searle JR (1995) The construction of social reality. The Free Press, New York
Simon HA (1996) The sciences of the artificial, 3rd edn. MIT Press, Cambridge, MA
Starreveld RW, de Mare B, Joels E (1994) Bestuurlijke Informatieverzorging (in Dutch), vol 1. Samsom, Alphen aan den Rijn
Suh B, Chi EH, Kittur A and Pendleton BA (2008) Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard. Paper presented at the proceeding of the twenty-sixth annual SIGCHI conference on human factors in computing systems, Florence
Turilli M, Floridi L (2009) The ethics of information transparency. Ethics Inf Technol 11(2):105–112. doi:10.1007/s10676-009-9187-9
Vaccaro A, Madsen P (2009) Corporate dynamic transparency: the new ICT-driven ethics? Ethics Inf Technol 11(2):113–122. doi:10.1007/s10676-009-9190-1
Van de Poel I (2011) The relation between forward-looking and backward-looking responsibility. In: Vincent N, Van de Poe I, Van den Hoven J (eds) Moral responsibility. Beyond free will and determinism. Springer, Berlin, pp 37–52

Vishwanath T, Kaufmann D (2001) Toward transparency: new approaches and their application to financial markets. World Bank Res Obs 16(1):41–57

Walton D (1996) Argument schemes for presumptive reasoning. Lawrence Erlbaum, Mahwah

Weber RH (2008) Transparency and the governance of the internet. Comp Law Secur Rev 24(4):342–348. doi:10.1016/j.clsr.2008.05.003

Westerman P (2009) Legal or non-legal reasoning: the problems of arguing about goals. Argumentation 24:211–226

Yu E (1997) Towards modelling and reasoning support for early-phase requirements engineering. In: Proceedings of the 3rd IEEE international symposium on requirements engineering (RE'1997), IEEE CS Press, pp 226–235

Zuiderwijk A, Janssen M (2014) Open data policies, their implementation and impact: a comparison framework. Gov Inf Q 31(1):17–29

# Design for the Values of Democracy and Justice

Auke Pols and Andreas Spahn

## Contents



**Abstract**

In this chapter, we provide an overview of literature on the relation between technology and design and the values of democracy and justice. We first explore how philosophy has traditionally conceptualized democracy and justice. We then examine general philosophical theories and arguments about this relation, dealing with the conception of technology as being "value-free" as well as with pessimistic and more optimistic assessments with regard to technology's potential for advancing democracy and justice. Next, we turn to three concrete design methods that seek to promote democracy and justice in the design process, namely, participatory design, technology assessment, and value-sensitive

A. Pols (✉) • A. Spahn
School of Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: a.j.k.pols@tue.nl; a.spahn@tue.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_13

design. Finally, we examine two cases of technology influencing democracy and justice: one regarding the relation between energy technology and democracy and one regarding the use of social media during the Arab Spring. We conclude that many pessimists focus on the "technological mind-set" as a problem that undermines democracy and justice; that in the absence of general design guidelines for democracy and justice, a focus on democracy and justice in the design *process* seems all the more important; and that design methods tend to include *values* rather than *theories* of democracy and justice, which suggests that a further integration of philosophy and the design sciences could create added value for both disciplines.

**Keywords**
Democracy • Equality • Justice • Non-neutrality of technology • Participatory design

## Introduction

From voting machines and solar panels for decentralized energy generation to the role of social media in helping to coordinate protests during the Arab Spring and spread information and footage over the globe, technology design has a clear impact on the values of democracy and justice. This impact has two main forms. First, the design of a particular technology can *weaken* or *strengthen* the values of democracy and justice, whether as an intentional choice or an unintended effect. *That* technology has this impact, however, does not mean that *designing* for democracy and justice is an easy job. Positive stories about communication technology spreading democracy and the "Facebook revolution" are balanced with cautionary tales about nondemocratic regimes using those same media for propaganda and tracking down protesters. Second, the values of democracy and justice can influence the design process itself, for example, through stakeholder consultation for inclusiveness or by arranging representation for stakeholders who do not have the power or capabilities to defend their own interests, such as very young children. This is not an easy job either, as a just and democratic process requires answers to thorny questions such as who exactly should be considered a stakeholder and whether the consultation process should aim for consensus or rather a compromise.

In this chapter, we provide an overview of the literature on the relation between technology/design and the values of democracy and justice. Particularly, in section "The Values of Democracy and Justice," we analyze the values of democracy and justice as they have been explicated in philosophy and present the main positions and fields of inquiry. In section "Topics in the Design for Democracy and Justice," we explore different philosophical analyses of the general relation between technology/design and democracy/justice. This section also discusses some of the mechanisms by which technology design can weaken or strengthen the values of democracy and justice. We analyze the conception of technology as being value neutral, as well as critical and more optimistic positions about the influence

technology has on democracy and justice. In section "Democratic Technology Design/Participation," we examine the philosophical ideas behind the practice of incorporating the values of democracy and justice in the design process and examine three design methods that seek to do exactly that: participatory design, technology assessment, and value-sensitive design. In section "Experiences and Examples," we elaborate two cases on the influence of technology on democracy and justice: one on energy production and networks and one on the role of social media during the Arab Spring. In section "Open Questions and Conclusions," we draw our conclusions and present a number of open questions for further research into designing for democracy and justice.

## The Values of Democracy and Justice

In this section, we analyze the values of democracy and justice from a philosophical perspective by indicating some of the key elements of these two values and by pointing out classical topics of the philosophical debates that circle around these two notions. The focus will obviously not be on presenting an exhaustive overview, but to highlight the aspects that are relevant in the context of technology design and its potential relation to questions of justice and democracy. We start by elaborating the notion of democracy before turning to justice. In sections "Topics in the Design for Democracy and Justice" and "Democratic Technology Design/Participation," we will look at the relation between technology and technology design and these two values.

### The Value of Democracy

Democracy is the ideal of a political system in which the citizens of a state are seen as essentially contributing to and determining the political power. Theories of democracies are as old as philosophy, although the modern understanding of democracy has its roots in Enlightenment ideals and later political philosophy. Already, Greek political theory distinguishes between six forms of states, depending on who is ruling ("one," "many," or "all") and whether the government strives for the greater good of all or departs from this ideal. These two criteria allow Aristotle not only to distinguish between tyranny and monarchy but also between *politie* – when the government of the people strives for the good – and *democracy*, which fails to meet the moral goal. Contemporary political theory is mainly interested in distinguishing different forms of government according to formal, value-free criteria, and the history of political philosophy can be read as a shift from moral evaluations to a separation of evaluative and descriptive claims (Hösle 2004). But next to sociological and political science approaches to democracy, philosophers are often concerned with normative democratic theory, i.e., the attempt to evaluate which elements belong to a democratic society and what the main arguments are in favor and against democratic structures of decision making (Christiano 1996).

In its broadest definition, democracy refers to a process of collective decision making, in which the members of the process have equality in participating and in which the decisions are made by the group and binding for all members of the group (Christiano 2008). In this sense, organizations just as well as states can be "democratic" to varying degrees. In political philosophy, a distinction is often made between direct versus representative democracy, depending on whether citizens cast a direct vote in a referendum on political issues or whether citizens elect representatives for a selected period. A central question in both forms of democracy is who counts as a citizen and has thus the right to vote, as well as the question how the election process is organized: equality, freedom, and secrecy are seen as crucial to guarantee a fair process (Hösle 2004, 526 ff).

Democracies are favored in many ethical theories as preferred organization of a state for both intrinsic and instrumental reasons. Among the instrumental reasons are arguments that point out positive consequences of a democratic form of government. In that sense, democracies are said to lead to better laws, as the government depends on the acceptance of the citizens for reelection and has thus an incentive to take their values, interests, and needs into account (e.g., Mill 1861; Sen 1999, p. 152). However, equality and majority votes as such do not guarantee fair outcomes, especially since majorities could systematically overrule minorities (Mill 1859). It has thus often been argued that democratic decision-making processes must be combined with solid protection of minority rights (Donovan and Bowler 1998; Hösle 2004).

Next to instrumental arguments in favor of democracy, philosophers have also defended democracy as having intrinsic value. It is a form of government that treats individuals as equal (Singer 1973, pp. 33–41) and is in line with the ideal of liberty (Christiano 2008) and autonomy of individuals. Especially Habermas has defended democratic decision making from an ethical perspective in which private autonomy and public collective legitimacy are linked closely together. Political decisions need the actual support of the affected citizens in order to be legitimate (Habermas 1994; Young 1990).

Both instrumental and intrinsic normative theories of democracy can be linked to the ideal of a just society, in which the decisions taken are fair and in line with the basic rights of the individuals, most importantly their general equality. Let us look therefore at the value of justice in turn.

## The Value of Justice

Humans are vulnerable and the things that may benefit or harm them, whether goods, opportunities. or emotional states, are not evenly distributed among all. Investigating which distributions are morally better than others is the domain of theories of *justice*. This section will provide a brief overview of philosophical discussions of justice and show several ways in which technology can be relevant for considerations of justice.

Discussions of justice can be grouped by the topic of their discussion. The two major topics of discussion are evaluating whether a given *state of affairs* is more

just than another, which is the subject of the field of *distributive justice*, and evaluating whether a *procedure or process* is more just than another, which is the subject of the field of *procedural justice*.

*Distributive justice* investigates distributions within states of affairs, where the main questions are (a) what exactly should be distributed (goods, opportunities, welfare, etc.), (b) who are the donors/receivers of those goods (individuals, groups, nations, etc.), and (c) what principle should determine the distribution (equality, desert, free trade, etc.). Technology can play an important role in both (a) and (b), as good to be distributed, as part of the system that facilitates distribution, and in applying distribution rules, e.g., through registering income tax or determining an individual's identity or age, which may be relevant in, for example, immigration procedures (Dijstelbloem and Meijer 2011).

Traditionally, philosophers working on distributive justice have focused on how nation-states should distribute goods or opportunities among citizens. The most important work in this tradition is John Rawls' *A Theory of Justice* (1971, 1999). Rawls proposes a basic distribution in two parts, where (1) everyone should have equal basic rights and liberties and (2) any social and economic inequalities may only result from a situation where there is fair equality of opportunity and when these inequalities are to the greatest benefit of the least advantaged members of society.

More recently, philosophers of justice have extended their research into distributions on an international level (*international* or *global justice*), which raises unique problems. Discussions here have also been inspired by Rawls (1971) and especially by his (1999) *The Law of Peoples*. Rawls, however, argues that there are no international structures that resemble the basic (democratic, liberal, legal, economical) structures of the nation-state sufficiently to create a demand for principles of justice on an international level. Rather, he focuses on the principles by which nation-states should set their international policies (cf. Nagel 2005; Freeman 2006).

Critics of Rawls have argued that his focus on nation-states misses the point here. They point out that globalization has given rise to many international systems that all affect the global distribution of burdens and benefits, such as UN institutions, trade regimes and embargoes, transnational corporations, NGOs, etc. They claim that the fact that our participation in these systems affects this global distribution itself raises obligations of justice, irrespective of how these systems are structured (Pogge 2002; Caney 2005; Cohen and Sabel 2006; Young 2006). It is important to note here that many of these systems are socio-*technical* and strongly rely on esp. monitoring, communication, and transport technologies as *means* for their day-to-day functioning. Other systems use technology as an *end* in pursuit of decreasing unjust distributions, e.g., international technology transfer or design for development (Oosterlaken 2009).

Finally, the immense power that technology has given us not only spans the globe but reaches far into the future as well. The way we deal with nuclear waste and climate change, for example, and how we shape our socioeconomic systems, have consequences for the rights and quality of life of future generations

(cf. Jonas 1979/1984). This raises demands of *intergenerational justice*: how should we distribute burdens and benefits over generations across time?

As with other questions of justice, John Rawls (1971) has opened the discussion on this, proposing a "just savings" principle. This requires currently living people to do two things: establish lasting and just institutions and save enough resources for future generations so that they will enjoy at least a minimum level of well-being. This makes Rawls' account *sufficientarian*, in that we should leave our descendants *sufficient* resources for a life worth living. Moreover, once just institutions are in place, saving sufficient resources should become "automatic." Unfortunately, realizing this is a major challenge, as many global institutions are not only unjust in space, as Pogge (2002) has noted, but also unjust in time, providing us with benefits (e.g., nuclear energy, fossil fuels) while passing the burdens (nuclear waste, climate change, and energy scarcity) on to future generations (Gardiner 2011).

International and intergenerational theories of justice are all branches of *distributive justice*, which evaluates states of affairs. In contrast, *procedural justice* is concerned with just *processes or procedures*. This idea underlies, for example, the democratic system, where governments are constituted not after abstract ethical ideals but according to the outcome of an election procedure following prespecified rules. Procedural justice is a fundamental part of *discourse ethics* (Habermas 1984, 1987) and *deliberative democracy* (Bohman and Rehg 1997). The basic idea is that a decision or policy is just and legitimate if it is the result of a public deliberation based on rational arguments. This implies that every participant should have the right to speak and be heard; no force is exerted except for that of the better argument, etc. (cf. Habermas 1990). Discussion should not only be on means and indicators, where goals are set in advance by the organizing stakeholder (such as the state). Rather, it should address both values/goals and indicators/means. A diversity of perspectives tends to be encouraged, as this increases the chance that new facts, conflicts, but also opportunities regarding the topic are brought into the discussion (Swierstra and Rip 2007).

Engineers and philosophers of technology tend to be quite interested in possibilities for making design procedurally just, and we will examine various proposed methods for this in section "Democratic Technology Design/Participation."

## Topics in the Design for Democracy and Justice

In this section, we want to first look at how the relation between modern technology and democracy and justice is analyzed in different contributions to the philosophy of technology. If one intends to strengthen the values of democracy and justice through design, one is well advised to first contemplate on the possible impacts of technology on democracy or political power structures as such. Many philosophers of technology – especially from the continental tradition – address issues of design for democracy in the context of a broader analysis of the relation between technology and society as such, while other thinkers – especially from the analytic tradition – focus more on concrete technologies and their relation to democracy and justice.

In principle, there are at least three possible positions, according to the ideas of neutrality and positive or negative influence: (i) technology could be seen as mostly neutral and value-free; (ii) it could be argued that certain technologies endanger democracy and justice; or (iii) that certain technologies promote and foster democracy and justice.

(i) Technology and democracy could be regarded as completely independent entities, such that issues of technology design and, e.g., social power structures do not interfere with each other. Even if social power structures would determine technology design (but not the other way around), it could be maintained that technology is value neutral. If in this way technology would be value neutral with regard to political aspects, such as power division and aspects of exercise of individual autonomy and issues of social distribution of resources and power, ethics of technology could be done for the most part without analyzing the underlying social structures of political power. Accordingly, a theory of society could for a large part either ignore technology or treat it as one social reality next to other social phenomena. In short, this position presupposes the idea that technology is largely apolitical and can be used in different social contexts. This position is still compatible with the idea that both modern technology and modern societies are expressions of an underlying common spirit (e.g., the same "episteme" in the terms of Foucault) or the result of similar processes (such as the process of "rationalization" in the terms of Max Weber). The key idea of the neutrality thesis is rather to deny that there is a strong relation of *causal* influence between the two phenomena of social order and technology. Therefore, both spheres – technology and democracy – can (and probably should) be studied separately from each other. This position emphasizes the ethical or at least political neutrality of technology.

The other two positions (ii and iii) assume that technology is not value neutral with regard to political issues. In his influential book *Critical Theory of Technology*, Feenberg distinguishes substantive and instrumental theories of technology (1991). According to instrumental views, technologies are neutral tools, which can be used in all different political and cultural circumstances without affecting or influencing the political order. According to substantive theories (such as Heidegger or Ellul), technology has an impact on social order and is politically not innocent. Positions that see technology (or certain technologies) either as a threat to democracy and justice or as a great promoter of justice and democracy thus often – though not necessarily – belong to the family of substantive theories of technology.

If technology affects important social values and leads to, or even presupposes, specific relations of power, then the relation between technology and democracy becomes more complex. For the sake of brevity, we will distinguish an optimistic and a pessimistic stance of this relation. The optimistic position would argue that modern technologies do often amplify democratic structures and have thus a positive effect on the implementation and flourishing of democracy. Especially, the Internet and new (social) media are often believed to have a positive impact on democracy (see section "Technology as an Amplifier of Democracy and Justice"). The pessimistic position would, however, argue that modern technologies often undermine, or at least endanger, social justice and democratic structures as they

contribute to the establishment of fixed power relations in which a technocratic elite is needed in order to control complex technological systems such as railroads and nuclear power plants (see section "The Critical Stance: Technology as a Threat to Democracy and Justice").

Even though the affirmative and the critical positions point in different directions, they do not necessarily contradict each other. Both positions assume that technologies are not value neutral (and, more specifically, not neutral with regard to social and political values). One could therefore argue that certain types of technologies – such as social media – are beneficial to democratic systems, whereas other types of technologies – such as nuclear power plants – require central control and therefore stand in tension with the central aspects of democracy. In the following, we will nevertheless present both positions separately for the sake of analytical clarity. We will present their main arguments and illustrate the consequences for the ethics of technology design. We will begin with the critical position before turning to the optimist's camp.

## The Critical Stance: Technology as a Threat to Democracy and Justice

If one puts the affirmative and the critical stance into the context of the recent history of philosophy of technology, one can trace back their roots to the difference between the enlightenment embracement of sciences and technology, as important elements of social progress, and the romantic discontent with radical changes within society that led to a loss of traditional values and cherished belief systems (Mitcham 1994, 275 ff.; Spahn 2010). The optimists – or the "modernists" – in this debate embrace the enlightenment ideals of reason and rational inquiry as a key to social and moral progress. They mainly regard modern science and scientific knowledge that is based on observation, experimentation, and mathematization as a powerful tool that the ancient world did not bring forth. As Mitcham has noted, this optimistic focus on the positive aspects of modern science and technology is a very natural position for engineers and accordingly a philosophy of technology that is close to the engineering perspective (Mitcham 1994, cfr.; Snow 1959). It is important to notice that the affirmative position – which will be discussed in more detail in the next section – often highlights moral reasons for the embracement of modern science and technology. Technological progress contributes to the taming of nature and frees humans from many calamities and burdens. Modern technology can contribute to the overcoming of poverty, help in the fight against diseases, free man from the hardship of labor, and contribute to many luxuries of modern life: an optimistic perspective of technology that we find as early as in Francis Bacon (1620) and that is still vivid in technology futurists such as Kurzweil (2005).

This optimism has, however, not been without criticism. More pessimistic views on technology can be found as early as antiquity (Mitcham 1994, p. 277) but can mainly be traced back to the philosophers that are skeptical with regard to the

project of modernity and its focus on science, such as Vico (1709) and Rousseau (1750). Mitcham calls this position *Romantic Uneasiness about Technology* which includes thinkers that reflect on the radical changes in society due to industrialization and continues up to the current time in philosophers who worry about the destructive potential of modern technologies, as reflected in nuclear bombs and environmental damages of overconsumption.

In the following, we give a brief overview of some of the most influential philosophical approaches that highlight the problematic aspects of technology. The main aim is not to go into bibliographical details, but rather to present the main arguments made about the negative influence of technology on democracy and justice. Let us look at some influential claims in the pessimist's camp.

1. Technology is more than a neutral tool. It is a form of approaching reality, that is in itself dubious and problematic and reduces nature to a "standing resource" and an object of manipulation.

   This view is most prominently developed in continental philosophy of technology by Martin Heidegger in his essay *The Question Concerning Technology* (1977/1953). For him, technology is not so much the realm of artifacts created by humans, but a way of disclosing reality under a very specific perspective. Heidegger rejects the instrumental view; according to which, technology is a human tool for reaching certain aims. Rather, technology is in essence a mind-set, a perspective of the world under which everything that exists is seen as a potential resource for manipulation. Technology discloses ("entbergen") reality in a very different form than, e.g., art does: it grasps reality under the perspective of usability for external purposes. Nature gets thus reduced to a standing resource ("bestand") for human activity – a very different way to approach nature from the one taken by an artist. Heidegger also opposes the anthropological view of technology; according to which, technology must be mainly understood as a human creation. He rather urges us to regard technology not so much as a human activity but as a force greater than us: human beings are not so much in control of technology, but driven by its powers. That poses a radical danger to humans, not primarily because of concrete risks of any given technology but more so due to the fact that we lose sight of other ways to approach nature. Everything, including humanity itself, gets converted into a mere object of manipulation.

   This idea, to link technology with a mind-set of control, manipulation, and domination, became a frequent topic in many different approaches of philosophy of technology and the relation between technology and democracy, and was taken up, among other domains, within the environmental ethics movement (cfr. DeLuca (2005) for an analysis of the relation between Heidegger's work and environmentalism). Hans Jonas (1979/1984), a pupil of Heidegger, describes the difference between ancient and late modern technologies as one of the key challenges of modern ethics. Whereas in previous time, technology was mainly a tool helping man to survive in a dangerous and hostile nature, it is now our task to protect nature from far-reaching, irreversible, and potentially disastrous

consequences of modern technologies. In a similar line, Ellul (1964/1954) analyzes the threat that modern technology poses to human freedom and broader humanity, pointing out the dangers of modern technology that strives for absolute efficiency in all human domains.

2. Modern technology leads to substantial changes in social order, benefiting a small elite at the expense of deskilling workers and creating social injustices

   While Heidegger, Jonas, Anders, and Ellul comment on the essence of technology and its impact on modern civilization, it is especially the Marxist tradition including the early and later critical theory that draws attention to the political and social justice implications of modern technology.

   Following Hegel, Marx links technology to the necessity of labor to sustain human life (Marx 1938). But his main focus is on the analysis of the immense social and political consequences of modern industrial labor. Modern mass production leads to a division of labor and a deskilling of the workforce as industrial technologies do no longer require skillful craftsmanship. This leads to an alienation from work, as little expertise is needed to work in factories. Following Marx, Feenberg argues thus that it is not enough to just focus on the ownership question of the means of production, but to rethink the design of technology itself, especially in the context of labor theory (Feenberg 2002). In a similar vein, Noble tries to show how certain design choices lead to deskilling and social injustice (Noble 1984).

   According to the Marxist tradition, these political implications are tremendous, creating great injustices in society. The means of production are possessed by a small elite that exploits the workforce, creating a class struggle for political power. As such, however, the Marxist theory is not a substantive theory of technology, as it is not technology in general that stands in opposition to a just distribution and democratic ideals, but the underlying capitalist economic system of distribution of power and economic resources. The influence of Marxist thinking on the political aspects of modern technology can hardly be overestimated, both in the political arena and in philosophical analysis. In this brief summary, we only focus on a few later contributions, mainly from the Frankfurt School and critical theory, before moving to the STS approach to technology, justice, and democracy.

3. Technology is rooted in the will to dominate nature and will inevitably lead to the domination of man as well, unless counter-measures are taken.

   Adorno and Horkheimer (1979) place modern technology in the greater context of enlightenment rationalism and political philosophy. They interpret modern rationality as a means-end rationality that aims at the control and domination of the external world. In essence, technology is a powerful tool to ensure human self-preservation against nature. The striking feature of technological-scientific rationality is the quantification of natural relations in order to make a controlled and predictable use of natural laws for the exploitation of nature for human means. This strategy of control and domination of the outer nature will inevitably also lead to a domination and degeneration of humans' inner nature. Since Adorno identifies modern enlightenment rationality

with this strategic tool of domination and exercise of power, a countering force can only come from outside of rationality. Much like Heidegger, Adorno seeks remedy in the arts, as a different approach to reality that is not inspired by the will to dominate (Adorno 1999).

Habermas builds forth on many insights of the early Frankfurt School. However, he argues for a more nuanced theory of rationality and thus a more nuanced approach of technology that comes closer to the instrumental theory of technology. He distinguishes between different types of rationalities, which are rooted in different basic anthropological interests and lie at the heart of different types of sciences (Habermas 1971). He tends to agree with Adorno that science and technology are mainly rooted in the will to dominate nature. They have their background in instrumental knowledge. But next to this impulse, there are other types of rationalities rooted in different anthropological features: the humanities are rooted in practical knowledge and aim at communication, understanding, and agreement on shared norms and values. The emancipatory interest finally strives at freeing oneself from dogmatic dependences and lies at the heart of psychoanalysis and rational criticism. Since these other perspectives (next to scientific and technological knowledge of nature) are also seen as types of rational discourse, it follows that a rational critique and guidance of technology is possible without referring to other sources of knowledge outside of rationality. Habermas' main contribution lies accordingly in the development of a theory of communicative rationality (Habermas 1987) and discourse ethics, which has inspired visions of participatory technology design or participatory technology assessment (e.g., Kowalski 2002, p. 14). In his theoretical works, Habermas defends the *lifeworld* (Lebenswelt) as a realm of communicative action, in which we discuss about and agree upon social values. In the world of modernity, the lifeworld is under constant threat of being invaded by the world of strategic rationality, a process that Habermas has coined the "colonization" of the lifeworld (Habermas 1987). This thesis mirrors the worry that we find also in Heidegger and Adorno and that has been discussed above: the concern that instrumental, technological rationality that centers on efficiency and strategic manipulation dominates more and more aspects of society that should not follow the logic of strategic rationality. This would mean for the values of democracy and justice that the realm of deliberation, which belongs to communicative rationality, is under constant threat to be replaced or invaded by the realm of strategic rationality. Democracy must thus always actively strive to defend itself against the pitfalls of technocracy and the ruling of a selected class of technocratic experts (Habermas: 1971; Feenberg 1992).

The most elaborated application of critical theory to technology has been presented by Andrew Feenberg (1991, 2002). He rejects both substantive and instrumental accounts of technology and regards critical theory as holding the middle ground between the two (2002, p. 5). Technology and the spreading of instrumental rationality is not a destiny that is beyond human intervention or repair. Critical theory thus wants to avoid the utopianism often associated with Marxist perspectives of technology and the resignation that Feenberg sees

present in views that claim that an alternative to the Western capitalist system of technology is not possible. Rather, he denies that "modernity is exemplified once and for all by our atomistic, authoritarian, consumerist culture. The choice of civilization is not decided by autonomous technology, but can be affected by human action" (ibid., p. 14). Currently, Feenberg sees the values of a specific social system and the interests of its ruling classes installed in the very design of current technology. But through radical democratization of technology design, a "shift in the locus of technical control" (ibid., p. 17) is possible that avoids problematic capitalist phenomena such as deskilling of labor and can thus help realizing suppressed human potentials.

The last view by Feenberg thus points already to the option that technology may indeed under favorable circumstances be a positive contributor to democracy and justice (see section "Technology as an Amplifier of Democracy and Justice"). According to Feenberg, technology design thus matters for the crucial impact it has on the division of power within society.

4. Technologies can be inherently political by settling an issue or being highly compatible (if not requiring) a given set of power.

Within STS, the point has often been made that the instrumentalist perspective on technology is mistaken and that technologies are in fact not morally neutral tools. According to Winner (1980), artifacts can be political in two ways. Either they can directly settle a political issue or the operation of certain technologies requires or at least suggests a certain structure of political organization. Artifacts can directly settle an issue by, e.g., excluding certain users from access or benefits of these technologies or by contributing to a redistribution of influence and power in favor of a capitalist elite. With regard to the second aspect, Winner has argued that certain technologies require a strict hierarchical and authoritative control in order to work: nuclear energy is one example that Winner discusses, whereas decentralized small solar cells can easily be combined with a more localized, decentralized, individual, and democratic control. In a similar vein, Shelley (2012) has argued that for monitoring technologies or risky technologies, issues of fairness can result from choosing the "prediction cutoff" point, which determines the relation between false positives and true negatives. She gives the example of the ShotSpotter, a system to detect and locate gunshots in urban areas. If the system is hypersensitive, it will give more false alarms, which will result in increased police presence and possible disruption of social life. If the system is not sensitive enough, public security may be compromised due to gunshots not being responded to. The risk here is that the dominant user or designer may establish the prediction cutoff point based on private interests rather than on more objective considerations of fairness.

Akrich (1992) analyzes the way in which the expected use context is inscribed into technologies, by what she calls the "script" of a technology. Like a theater play, a technology designer "in-scribes" a vision about how the technology is supposed to be used that pre-scribes the user how to deal with it. Users can of course creatively ignore this script ("de-scription"), but nevertheless, scripts in technologies always distribute and delegate responsibilities

(cfr. section "Energy Production, Justice, and Democracy" for an application of these ideas to energy technologies).

In line with this, the actor-network theory (Latour 1979, 2005; Law 1999) emphasizes that technologies have "agency" and influence the network. Scholars in STS have thus tried to identify the role of power differences and power distribution in actor-networks that in part are reinforced or materialized in given technologies. In order to avoid stabilizing one-sided hierarchical power relations and social injustices, the argument has been made that technology design needs to make sure to include neglected interests of marginalized groups. Participatory technology design thus tries to include the values and interests of all affected parties prior to developing and implementing major new technological systems (see section "Democratic Technology Design/Participation").

## Technology as an Amplifier of Democracy and Justice

As we have seen, there are many attempts to account for the tension or even negative impact of technology on democracy, justice and social power relations. However, many authors argue that we can be more optimistic about the relation between technology and democracy and justice. They argue for the potential of technology to contribute to those values, though often not without reservations. In this section we will consider various arguments made why technology can or will further democracy and justice and critically examine them.

1. Technology contributes to human welfare and capabilities.

   In order to be able to participate successfully in a democratic society, citizens need to have their basic needs met with regard to food, water, shelter, etc. Meeting these basic needs, or a certain level of well-being, is also a demand of justice under a sufficientarian conception. Technology can help increase well-being, keep us safe from harm, etc. Thus, it can help fulfill the conditions for democracy and justice.

   The argument that technology contributes to human welfare is the oldest and perhaps the least controversial argument for technology as enhancing democracy and justice. Its best-known historical proponent is Francis Bacon (1620, 1627; cf. Mitcham 1990), who extolled the virtues of modern technology in conquering nature and bringing about social and moral progress. Though this optimism was tempered later, partly in reaction to the negative aspects of industrialization (see previous section), many engineers still subscribe to it. In Winner's words, "The factory system, automobile, telephone, radio, television, space program, and of course nuclear power have all at one time or another been described as democratizing, liberating forces." (1986, p. 19). Those who take this argument to its furthest extreme argue that our technology will someday turn us into super- or post-humans, helping us to fully overcome our biological limitations and making considerations of justice obsolete (Kurzweil 2005; Savulescu and Bostrom 2009).

*Criticism against this argument* takes two main forms. First, it can be argued that technology *can* contribute to well-being, but does not necessarily have to do so: it can also diminish it through direct or indirect harm, new risks, exploitation of humans and nature, etc. Second, it can be argued that even if technology contributes to welfare, this is no guarantee that it will contribute to democracy and justice. For example, if a technology were to contribute to a massive increase of welfare for the rich but only a marginal increase of welfare for the poor, it would not be just according to Rawls, as it would violate his condition that any inequalities in distribution should deliver most benefits to those who are worst off (see section "The Value of Justice").

2. Technology can enhance one's skills and knowledge regarding how to participate in a democratic system / further justice.

   Having one's basic needs met can be regarded as a condition for democratic participation but so is knowing what is going on and knowing how to participate. With regard to the knowledge of what is going on, technology has greatly advanced methods of communicating and sharing information, from the advent of the printing press to newspapers and television and later information and communication technology (ICT) and social media. With regard to knowing how to participate, technology has contributed to education and the dissemination of information about procedures, contact details of civil servants, etc.

   *Criticism against this argument* is very similar to that against the previous argument. Johnson has argued that our global information infrastructure does not automatically empower the people. She argues that more power does not simply come through *more* information, but through *accurate*, *reliable*, and *relevant* information (1997, p. 25). Indeed, without any filter, the deluge of information we get through ICT can easily distract rather than empower us (Floridi 2005), and propaganda spread by authoritarian regimes, as well as entertainment to placate the masses, may hamper meaningful democratic engagement. Consequently, the democratic potential of ICT depends very much on who filters the information and what criteria they use. See, for example, Massa and Avesani (2007) on how different kinds of trust metrics, ways to rate how trustworthy a participant in an online community is, can introduce different biases in the information exchanged in those communities.

   A broader problem behind this is the practical issue that even engaged citizens are limited in the time and resources they can invest in gathering and judging information and discussing and taking action on the basis of that information (Bimber 1998; Van den Hoven 2005). Thus, more information and possibilities for digital activism *cannot* even increase political activity above a certain level unless this "information cost" is brought down, making unbiased design of filters and facilitating discussion tools all the more important. Van den Hoven argues that this problem necessitates a rethinking of what our democracy should look like. Rather than striving for the unrealistic ideal of the Well-informed Citizen, he argues, we should aim at the more practical Monitoring Citizen (Schudson 1998). The Monitoring Citizen does not *know* everything that is going on but can *monitor* it successfully and can investigate and contest policy when needed.

3. Technology can facilitate decentralised communication and coordinated action.

   Technology cannot only help us decide what to do but also to actually do it. This can advance democracy and justice by giving more power to the people even in nondemocratic states. For example, during the Arab Spring, social media were widely used to coordinate protests even where official media channels were blocked by authoritarian regimes (see section "Arab Spring/ICT").

   *Criticism against this argument* is that technology does not automatically *increase* communication and coordinated action; it may rather *change* communication and action patterns, connecting some social groups while isolating others. Johnson (1997) argues that in the past, shared geographic space determined shared contact and action. With our global ICT network, she warns that individuals may more and more contact and work together with like-minded individuals rather than with others with diverse and conflicting viewpoints, which may lead to less involvement in local or national communities and ideological isolation. This warning is echoed by Sunstein (2001), who advocates (among other things) for new public electronic forums where discussions are encouraged between people of different viewpoints. Sclove (1995) is similarly ambivalent, seeing new options for cooperation, such as the creation of "virtual commons," but also risks of disintegration of local social groups. Furthermore, new technologies such as social media may offer opportunities for distributed communication, but these may be partly or fully offset by the opportunities it offers powerful actors such as states for monitoring and controlling this communication (Morozov 2011). Pariser (2011) links this ideological isolation and propaganda spreading to the workings of the algorithms that operate behind the scenes of Internet giants such as Google and Facebook. He warns that these algorithms, which most people do not know and do not think about, have a great influence on what we see and hear on the Internet, creating a "filter bubble" where we only get to see information that corresponds with our own views (personalized information) or those of the state or the company (propaganda). Then again, others have suggested that personalized information may expand rather than constrain our views, which could further the values of democracy and justice (Hosanagar et al. 2014). For the relation between search engines and democracy, see also Introna and Nissenbaum (2000), Nagenborg (2005), and Tavani (2014).
4. Technology may draw previously disinterested parties into democratic processes.

   In presenting her ideal of "collaborative democracy," Noveck (2009) argues that technology can draw people into democratic processes who would normally not participate, e.g., by offering alternative procedures to classical deliberation and enabling people to enter those processes themselves rather than having to be selected as "experts" by policy makers who may be biased in their selection procedure.

   *Criticism against this argument* could be that the threshold for democratic participation is not necessarily lowered, but rather changed: interested parties still need an Internet connection and the know-how to join and participate in digital forums. Furthermore, even if technology can lower the overall threshold

for participation, some voices cannot be drawn in, such as those of disinterested parties or future generations who will be affected by the decisions that are taken now (Thompson 2010).

5. Technology does not discriminate according to human biases.

   It is a well-documented fact that humans suffer from psychological biases that may lead to discrimination and injustice in decision making (Sutherland 1992/2007). Technology does not suffer from these biases: as Latour puts it, "no human is as relentlessly moral as a machine" (1992, p. 157).

   *Criticism against this argument*, however, is also provided by Latour, who argues that technology contains scripts for prescribing behavior that can be inscribed by engineers both intentionally and unintentionally, and this prescription can lead to different forms of discrimination (cf. Winner 1980 on the architect Robert Moses' "racist overpasses"). To borrow an example from Latour, hydraulic door closers will not discriminate based on gender or race, but they will discriminate against those who are too weak to push the door open, such as the young or the elderly. Introna (2005) gives the example of gender/race biases in face recognition systems and proposes "disclosive ethics" as a way to deal with biases generated by technologies that are relatively closed to scrutiny. Friedman and Nissenbaum (1996) give examples of biases in computer systems and offer several suggestions on how to identify and deal with them. Roth (1994) argues that complex ballot design may bias democratic elections by excluding or misleading undereducated voters. All in all, technology may not suffer from the same biases in the same way that humans do, but bias can still enter design in many ways and so lead to undemocratic or unjust situations.

To summarize, while technology certainly has the *potential* to contribute to democracy and justice, it can often also easily be used for its opposite. Various design approaches have attempted to utilize this positive potential and avoid its pitfalls by introducing democracy and justice in the design *process*, in the hope that this leads to a more democratic and just *product*. We will examine some of those approaches closer in the next section.

## Democratic Technology Design/Participation

Democratic technology design starts from the assumption that technology has a major impact on human life, but has in the past often not adequately included the values of democracy and justice in the design process itself. As discussed in the introduction, one of the arguments for democratic decision making is that the affected persons need to have a say in the decision-making process in order to make sure that their values and interests are taken into account. Democratic technology designs or visions of participatory technology assessment start from this idea and move it from the legal domain to the technological domain. Since technology is a major factor, all sides should be heard in the design and/or the implementation of important technological projects.

The ideas of participatory processes of decision making are a natural part of political theories that highlight democratic elements. With regard to ethical theories, visions of participatory design often go back to contractualism (Scanlon 1998) or discourse ethics (Apel 1973; Habermas 1993). Both ethical approaches regard the idea of a consensus that no one can reasonably reject or that every affected party should be able to accept, as a key ideal of normative theory. As Scanlon argues, "An act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behaviour that no one could reasonably reject as a basis for informed, unforced, general agreement." (Scanlon 1998, p. 153). In a similar line, Apel and Habermas aim at formulating the criteria for the ideal discursive community that should guide real processes of democratic decision making. Among these are a power-free dialogue, consistency, transparency, and a striving for common interest (Habermas 1993). It has been objected that a consensus is often not possible, so that real participation should not always be consensus oriented. Rather, it should follow the negotiation model of compromise-oriented discussions to be relevant for practice (e.g., Van den Hove 2006). Next to the question whether the ideal of agreement or compromise should guide participatory processes, philosophers have analyzed the question at which points participants may even have a moral right to leave the deliberative process and take measures of political activism, especially in the absence of equality of the participants and/or severe divergence from orientation at the ideal speech situation by some or all of the participants (e.g., Fung 2005).

In the remainder of this section, we therefore discuss a number of concrete methods that have been proposed to bring principles of democracy and justice in the design process: participatory design, constructive technology assessment, and value-sensitive design. Before we start, some caveats are in order. First, let us note that there is a difference between design *for* all and design *with* all: universal or inclusive design, extending the potential user group as much as possible (Clarkson et al. 2003), can be very good and useful, but it is not a blueprint for a democratic or just design *process* (though it does not exclude such processes either). Second, some concepts such as that of *responsible innovation* (Stilgoe et al. 2013) are ambiguous in the sense that they can be constrained to artifact design but also include innovation trajectories, institutional change, and even the decision to abandon proposed innovations that turn out to be deeply problematic in some way, thus not leading to the design of new technical artifacts at all. Third, while "stakeholder involvement" seems to be the mantra for many design methods that aim to be just or democratic, it is important to note that stakeholder involvement itself does not automatically make a design process just (Greenwood 2007). Unless great care is taken that the conditions for a just process are met (only and serious discussion of rational arguments, compensating for power differences, etc.), proclaimed stakeholder involvement can easily be abused to legitimize predetermined courses of action, as the following quote illustrates:

> In all these ways, European democracy is biotechnologized. Participatory exercises help legitimize the neo-liberal framework of risk-benefit analysis, which offers us a free

> consumer choice to buy safe genetic fixes. (. . .) If we wish to democratize technology, I suggest that we must challenge the prevalent forms of both technology and democracy. (Levidow 1998, p. 223)

*Participatory design* as a research field arose in the 1970s and 1980s when ICT entered the workplace. Early introduction of ICT was regularly met with hostilities from workers who felt that their interests were not adequately taken into account in this transition. This sparked research into how the needs and interests of workers could be incorporated in these management-driven transitions (Kensing and Blomberg 1998). Three main issues in the participatory design-literature have been (1) the politics of design, as it was quickly recognized that new technologies often supported existing power structures and management strategies of, e.g., centralized control, making it difficult to adapt them to workers' interests and needs; (2) the nature of participation, regarding when, how, and why workers should participate (e.g., Clement and Van den Besselaar 1993); and (3) developing methods, tools, and techniques for actually carrying out the participatory design process, such as the Cooperative Experimental Systems Development, which includes cooperative prototyping and design for tailorability (Grønbæk et al. 1997), and MUST, which focuses on creating visions for change for both the technology to be adopted and the organization that intends to adopt it (Kensing et al. 1998).

Despite the apparently ethical motivations and the fact that many recommendations of participatory design are in line with prescriptions of procedural justice, ethical reflection on participatory design principles is fairly recent and mostly concerned with explicating ethical issues rather than prescribing particular courses of action. Steen (2011) connects participatory design with ethics of the other, pragmatist ethics, and virtue ethics. Regarding ethics of the other, he identifies a tension between the need to be open to others (e.g., the users) and the need to, at some point, close the discussion and finish the design. This connects to the more general question in procedural justice of who should have the authority to establish that a workable consensus has been reached: ideally, this conclusion is established by everyone involved, but constraints of time and resources as well as the possibility of fundamental disagreements can make full consensus unfeasible or impossible. Regarding pragmatist ethics, Steen draws on Dewey's (1920) prescriptions for processes of inquiries, and regarding virtue ethics, he argues that it can help to reflect on the kind of person a designer should be in her or his role as discussion participant.

Robertson and Wagner (2013) explicitly explore the relations between ethics in general and participatory design, including the effects of design on the world we live in, difficulties in aligning ethical principles with politics in practice, dealing with value conflicts (e.g., whether severely ill children should be drawn into the participatory process or be protected from the strenuous task), and accounting for cultural differences in the participation process.

*Technology assessment* started out as a method to predict societal and ethical impacts of new technologies in an early stage. While this in itself does not make it a democratic or just method, it has been developed in accordance with these values into, e.g., *participatory technology assessment*, where many societal parties are involved in the ethical evaluation (Kowalski 2002), and *constructive technology*

*assessment*, where the results of the assessments are used not so much for regulatory purposes, but more for the redesign of the technology (Schot and Rip 1997). As with participatory design, however, while constructive technology assessment is based in part on the *values* of democracy and justice, it does not make the use of explicit *theories* of democracy and justice. This criticism is also given by Palm and Hansson (2006), who propose *ethical technology assessment* as an alternative. Palm and Hansson explicitly avoid commitment to one particular ethical theory, but they identify a number of ethical aspects of new technologies that should be taken into account. Democratic aspects include the dissemination of information about and control over new technologies; justice aspects include biases of gender, race, sexuality, etc., and biases against handicapped people, such as the cochlear implant that is perceived by some as a threat to the deaf culture. Some authors also argue that keeping technology flexible and open for multiple uses and reconfiguration is part of design for democracy (e.g., Van der Velden 2009; Kiran 2012; Dechesne et al. 2013).

Another method that explicitly seeks to involve stakeholders and their values into the design process is *value-sensitive design* (VSD; Friedman 1996; Friedman et al. 2005). Value-sensitive design employs a tripartite methodology, consisting of conceptual investigations into who is affected and what values are implicated by a new technology, empirical investigations into stakeholder values and trade-offs and the use context, and technical investigations into how a particular technology can help or hinder certain values. Though VSD does also not explicitly draw on theories of democracy and justice, concern with those values is witnessed by the fact that it seeks to include both direct and indirect stakeholders: not only the users of a technology but all those affected by it in some way.

Ethical criticism against VSD is that, while it does argue for the inclusion of direct and indirect stakeholders, it does not offer any methodology for determining who can be legitimately considered a stakeholder (Manders-Huits 2011). Also, it has been argued that VSD does not explicitly support a legitimate deliberative procedure for discussing stakeholder input and justifying trade-offs and that procedural justice theories, and particularly discourse ethics, can help in this regard (Yetim 2011). To summarize, while design is more and more oriented at the *values* of democracy and justice, the integration of the *theories* of democracy and justice into design methods and practices has only just started and still requires much work. We will examine some of the challenges for this integration in section "Open Questions and Conclusions."

## Experiences and Examples

In this section, we discuss two short cases. We intend to contrast a classical field of debate about the design for democracy and justice with a more recent discussion. We thus first look at the discussion about nuclear or solar energy and the effect of these choices on social power distribution and political institutions. After that, we investigate the recent debate about the power of modern mass communication and social media to promote democracy and undermine authoritarian power.

## Energy Production, Justice, and Democracy

Technologies for energy production have been at the center of a debate about "inherently political artifacts" (Winner 1980). In this section, we like to illustrate the broader claims made in the previous section by sketching the debate about the politics of energy technology. If one accepts the premise of inherently political technologies, one can ask in how far energy technologies are political.

The first consideration is that modern lifestyle depends on energy supply. The energy demand of modern civilization is one of the major drivers of international conflicts, especially given the dependence of the West on fossil fuels. Similarly, the threats of climate change are a result of an ever growing worldwide energy demand. Already Jonas (1979/1984) has thus argued, as we have seen above, that modern technology raises radically new questions with regard to responsibility and justice. How can we protect nature and how can we strive for a just distribution of resource consumption? On the one side, it has been argued that participation and civic engagement are essential in the implementation of major energy projects (cfr. Ornetzeder and Rohracher 2006; Lewis and Wiser 2007), but on the other hand, future generations by definition cannot participate in the process, as they do not yet exist. This has led to the question of how to represent future generations in policy making, especially in the case of sustainability and long-term-planning (Hösle 1994; Gosseries 2008).

With regard to democracy, however, another aspect next to participation becomes relevant in the analysis of energy technologies. It has been claimed that different types of energy production suggest different types of political institutions. As Winner (1980) has argued, nuclear energy requires a centralized top-down system of political control, whereas solar energy lends itself more easily to more decentralized individual bottom-up initiatives. In a similar line, Akrich (1992) has analyzed the impact electrification has on traditional society in developing countries and how different technologies redefine social roles. She argues that different types of energy technologies also define social roles and thus contribute to or emphasize differences in power structures. A power generator that can be used in rural villages strongly suggests a specific method to divide costs. The investment cost for the generator and the costs of using the generator (the fuel) can be easily separated, thus suggesting a whole microcosmos of economic relations between the "owner" and a "lender." Battery-driven lighting kits alike are not only designed with certain technical functional requirements in place but also include assumptions about the knowledge of the user, about maintenance structures and use context. Finally, the whole process of introducing electrification is likewise not only a technology transfer but imposes at the same time a system needed for payment, a legal structure of property rights and landownership, control of consumption and payments, and the like. The analysis that Akrich is giving is thus meant to illustrate that technologies go beyond fulfilling their function but also alter economic and political structures. The impact of energy technologies on social structures, justice, and civic engagement has since been analyzed in detail by philosophers and social science scholars (e.g., Chess and Purcell 1999; Devine-Wright 2013; Hoffman and

High-Pippert 2005). One illustrative case study by Nieusma and Riley (2010) shows that engineering for development initiatives that are explicitly aimed at social justice and implemented with careful consideration of the nontechnical aspects of decentralized small-scale energy technologies nevertheless often face serious challenges. One of the cases they present is the rural electrification campaign in the southwestern district of Monaragala by the Energy Forum in Sri Lanka during 2000–2002. The aims of the Energy Forum in Sri Lanka are the promotion and implementation of renewable technologies, focusing on decentralized energy technologies, including dendro power for rural electrification. Even though rural electrifications contribute to technological advancements, one of the central motivations of the Energy Forum is one of social justice: "How to more fairly distribute the nation's energy resources [. . .] so that the rural poor will benefit as well" (ibid., p. 43). The implementation also aimed at strengthening civic community and included nurturing "effective working relationships with village leadership and community members, [. . .] [and] two participatory design workshops where villagers shared their perspectives on the project" (Nieusma and Riley, p. 44). This is in line with the idea that decentralized, locally owned government energy projects foster social community and may even strengthen democracy (Hoffman and High-Pippert 2005). However, Nieusma and Riley conclude that despite the "purity of its motives and regardless of the effort put into transferring control in a sensible way" (ibid., p. 50), the project failed to empower the community members in the targeted village. They argue that designing for social justice requires overcoming engineering project risks such as (1) overfocusing on technology, (2) the occlusion of power imbalances in social interaction, and (3) ignoring the larger structural (including social) context (ibid. 51 ff.).

## Arab Spring/ICT

ICT and social media have been heralded as saviors as well as threats to democracy. Winner (1992) has already pointed out that the fax machine had helped revolutionaries to hasten the demise of the USSR, while Chinese revolutionaries using fax machines during the 1989 Tiananmen Square protests were quickly tracked down and arrested by the authorities. This section takes a closer look at the role of social media in more recent events, particularly those known collectively as the Arab Spring.

The Arab Spring has been said to have started by the self-immolation of the Tunisian fruit vendor Bouazizi, which quickly led to public protests and eventually the flight of the Tunisian president Ben Ali. The protests spread to other North African and Arabian countries, and governments were overthrown in Egypt and Libya. Whether these revolutions will in the long run lead to more democratic governments is still an open question, with the ongoing civil war in Syria and a planned mass execution of more than 500 civilians in Egypt at the time of this writing. However, it is undeniable that social media played an important role during these revolutions.

While the Arab Spring has been called a "Facebook/Twitter revolution," most academics seem to agree that social media were a necessary but not sufficient part of the uprisings. Khondker (2011) argues that two factors were crucial: the presence of revolutionary conditions and the inability of the state to control the revolutionary upsurge. Revolutionary conditions include high income inequality, government corruption, and high youth unemployment – a large percentage of Tunisia's and Egypt's populations are young, and many of those young people are unemployed, tech-savvy, and have no job or family responsibilities that would stand in the way of a willingness to participate in the protests (Howard et al. 2011; Lim 2012). In many ways, the Arab Spring uprisings were a culmination of social unrest, demonstrations, and social media protests that had been simmering for years (ibid.). An example of the inability of the state to control the upsurge is the Egyptian president Mubarak's shutdown of the telecommunications network, which was partly evaded by protesters with satellite phones and hindered government agencies and ordinary citizens as well (Howard et al. 2011). In the absence of the aforementioned revolutionary conditions, however, social media protests may well be stamped out by the authorities arresting protesters, infiltrating networks, and spreading propaganda.

A specific contribution of social media to the protests that has been mentioned is *the creation of spaces and networks* for connecting people with common interests (Allagui and Kuebler 2011). This option has already been mentioned by Sclove (1995). Johnson (1997) has warned against the opportunities the Internet offers for like-minded people to contact and agree with each other rather than get exposed to divergent opinions and worldviews, but the Arab Spring showed that networking did also occur across ideological and religious boundaries on the basis of shared complaints against the regime (Lim 2012). Moreover, rather than just leading to "slacktivism" and Facebook rants (Morozov 2011), the Arab Spring also showed that *injustices could mobilize those networks* to actually take to the streets and helped protesters to *coordinate* when they did. This happened, for example, after the brutal police murder in 2010 of Khaled Said, a young Egyptian claimed to have been targeted because he possessed evidence on police corruption (Lim 2012). Finally, social media have been used as an *alternative source of information* in countries with strong media censoring and, conversely, to *connect with traditional media* in other countries in order to get information and news out into the world (Khondker 2011).

Many philosophers of technology have proclaimed that technology is not "morally neutral." However, the role of social media during the Arab Spring has shown that it may well be neutral *with respect to certain values*: the openness and accessibility of information at least make it a valuable tool for both nondemocratic regimes and their protesters. There certainly are attempts to further "democratize" social media, for example, through participation in the Global Network Initiative, which has developed principles and guidelines for ICT companies for safeguarding user access, privacy, and freedom of expression (as of the time of this writing, Twitter is notoriously absent). However, there are attempts to bring social media

and other software under state control as well, e.g., by US engineers being required by the NSA to build in "back doors" for snooping. Security technologist Bruce Schneier has criticized this practice and issued an open call for software engineers to expose those practices and actively "take back the Internet" (Schneier 2013a, b). Clearly, there is more technical and ethical work to be done yet if social media are to be made (and kept) a truly democratic technology.

## Open Questions and Conclusions

In this section, we will identify three recurring themes and a number of open questions for both engineers and philosophers interested in furthering design for democracy and justice.

First, a general theme among the technology pessimists was the threat to democracy and justice of the "technological mind-set" that Mitcham (1994) has noted is a natural perspective for engineers to take. While in itself nothing is wrong with the values of effectiveness and efficiency or even with viewing nature as a "standing resource," this becomes problematic in excess. This occurs, for example, when people become unable to view nature in other ways (Heidegger), when other values or viewpoints are derided as "lesser" or "irrational," or when the technocrat mind-set is applied to areas where it should not be applied to (Habermas' colonization of the lifeworld).

Interestingly, while Heidegger and Adorno both identify this as a grave problem, they seek the solution outside of the domain that has created it, in the arts. While the arts can certainly make us look at the world in a different way, this is of little consolation to engineers who would like to design for democracy and justice. However, engineers interested in curbing the excesses of the technological mind-set could draw on work by Feenberg and Habermas, particularly the realization that other viewpoints represent different values and rationalities rather than simply being irrational and thus require serious consideration.

Second, while the technology pessimists discussed above at times tend to be overall pessimistic about the influence of technology on democracy and justice, most optimists tend to be nuanced, arguing that technology has great potential to advance those values – but also great potential to hinder them. Overall, there seems to be no design rule (yet) that, when applied to a technology, will make it (more) democratic and just in itself. Some factors that determine a technology's impact on democracy and justice may be technological. Many factors, however, are outside the control of the engineer, such as the willingness of the Tunisian populace to revolt during the Arab Spring. Other factors might be only under limited control of engineers, such as those that lie in the realm of use and institutional contexts. It is no wonder, then, that design methods that seek to further democracy and justice tend to focus on what engineers do have control over (though not necessarily full control): the design process. The third conclusion will therefore be devoted to this process and the relevant open questions for research that have been identified in this paper.

Third, in section "Democratic Technology Design/Participation," we have examined various design methods that strive to include the values, and sometimes theories, of democracy and justice. In treating these methods, a number of ethical questions cropped up again and again, progress on which is necessary and instrumental both for just and democratic design methods and for philosophical theories of democracy and justice. Those are:

1. *Who are the stakeholders*? Ideally, all those who "have a stake" in the design should be consulted (with possibly certain exceptions, such as business competitors and criminals). However, especially for novel technologies not (yet) embedded in a clear use context, such as nanotechnology, it may be impossible to predict beforehand who will eventually be affected by the technology. Moreover, potential stakeholders may include those who are not able to participate in the discussion, such as future generations and animals, and it may not always be clear who could and should legitimately represent their interests.
2. *Whose opinions should be taken into account?* This question has practical aspects, e.g., should severely ill children be dragged into a consensus procedure involving technology that will affect them (Robertson and Wagner 2013)? It also has theoretical aspects, however, e.g., whether people with extremely divergent, uncommon, or incoherent worldviews should participate in the procedure (Taebi et al. 2014). Generally, a diversity of views is seen as beneficial to the process as it may open up new possibilities. Additionally, excluding critical or unorthodox voices can easily become a tool of powerful participants to silence parties with opposing interests or an excuse not to involve minority groups or local communities (Levidow 1998; Van der Velden 2009). The greater the diversity of viewpoints and values is, however, the more difficult it may become to achieve consensus or closure.
3. *How do different opinions weigh?* The classic democratic tenet is "One person, one vote." However, this assumes that all stakeholders are affected equally by a particular decision or design, which may not always be the case. VSD, for example, distinguishes between direct and indirect stakeholders, but does not elaborate on what this implies for the weighing of interests.
4. *In case a consensus is not reached, who should make trade-offs and achieve closure?* Ideally, consensus is reached and trade-offs are made by mutual agreement. However, persistent disagreements and constraints of time and resources may make this impossible (Steen 2011). Giving particular groups (or particular ethical theories) the authority to achieve closure of a discussion may then be necessary, even though it runs counter to the prescriptions of procedural justice. Behind this practical problem lies thus a deeper philosophical problem that requires further investigation.

**Acknowledgments** This research was supported the MVI research programme "Biofuels: sustainable innovation or gold rush?", financed by the Netherlands Organisation for Scientific Research (NWO).

## Cross-References

- ▶ Design for the Value of Inclusiveness
- ▶ Design for the Value of Presence
- ▶ Design for Values in Nuclear Technology
- ▶ Human Capabilities in Design for Values
- ▶ Participatory Design and Design for Values
- ▶ Technology Assessment and Design for Values
- ▶ Value Sensitive Design: Applications, Adaptations, and Critiques

## References

Adorno Th (1999) Aesthetic theory, New edn. Athlone, London
Adorno TW, Horkheimer M (1979) Dialectic of enlightenment, New edn.Verso, London
Akrich M (1992) The de-scription of technical objects. In: Bijker WE and Law J (eds) Shaping technology/building society. MIT Press, Cambridge, MA, pp 205–224
Allagui I, Kuebler J (2011) The Arab Spring and the role of ICTs: editorial introduction. Int J Commun 5:1435–1442
Apel KO (1973) Transformation der Philosophie: Sprachanalytik, Semantik, Hermeneutik. Das Apriori der Kommunikationsgemeinschaft. Suhrkamp, Frankfurt a. M.
Bacon F (1620) The new organon. Cambridge University Press, Cambridge/New York
Bacon F (1627) New Atlantis: a worke vnfinished. In: Bacon F (ed) Sylva sylvarum: or a naturall historie, in ten centuries. William Lee, London
Bimber B (1998) The internet and political transformation: populism, community and accelerated pluralism. Polity 31(1):133–160
Bohman J, Rehg W (eds) (1997) Deliberative democracy: essays on reason and politics. MIT Press, Cambridge, MA
Caney S (2005) Justice beyond borders. Oxford University Press, Oxford
Chess C, Purcell K (1999) Public participation and the environment: do we know what works? Environ Sci Technol 33(16):2685–2692
Christiano T (1996) The rule of the many: fundamental issues in democratic theory. Westview Press, Boulder
Christiano T (2008) Democracy. In: Zalta EN (ed) The Stanford encyclopedia of philosophy, Fall 2008 Edition. http://plato.stanford.edu/archives/fall2008/entries/democracy/. Accessed 30 Mar 2014
Clarkson PJ, Coleman R, Keates S, Lebbon C (2003) Inclusive design: design for the whole population. Springer, London
Clement A, Besselaar P van den (1993) A retrospective look at PD projects. In: Muller M, Kuhn S (eds) Participatory design: special issue of the communications of the ACM, vol 36, no 4. pp 29–39
Cohen J, Sabel C (2006) Extram republicam nulla justitia? Philos Public Aff 34(2):147–175
Dechesne F, Warnier M, Van den Hoven J (2013) Ethical requirements for reconfigurable sensor technology: a challenge for value sensitive design. Ethics Inf Technol 15(3): 173–181
Deluca KM (2005) Thinking with Heidegger: rethinking environmental theory and practice. Ethics Environ 10(1):67–87
Devine-Wright P (ed) (2013) Renewable energy and the public: from NIMBY to participation. Routledge, London/Washington DC
Dewey J (1920) Reconstruction in philosophy. Henry Holt, New York

Dijstelbloem H, Meijer A (2011) Migration and the new technological borders of Europe. Palgrave Macmillan, Basingstoke
Donovan T, Bowler S (1998) Direct democracy and minority rights: an extension. Am J Polit Sci 42:1020–1024
Ellul J (1964) The technological society. Knopf, New York (french orig. 1954)
Feenberg A (1991) Critical theory of technology. Oxford University Press, Oxford/New York
Feenberg A (1992) Subversive rationalization: technology, power, and democracy. Inquiry 35(3–4):301–322
Feenberg A (2002) Transforming technology. A critical theory revisited. Oxford University Press, Oxford/New York
Floridi L (2005) Information technologies and the tragedy of the good will. Ethics Inf Technol 8:253–262
Freeman S (2006) The law of peoples, social cooperation, human rights, and distributive justice. Soc Philos Policy 23(1):29–68
Friedman B (1996) Value-sensitive design. Interactions 3(6):16–23
Friedman B, Nissenbaum H (1996) Bias in computer systems. ACM Trans Inf Syst 14(3):1–18
Friedman B, Kahn PH Jr, Borning A (2005) Value sensitive design and information systems. In: Zhang P, Galletta D (eds) Human-computer interaction in management information systems. M.E. Sharp, New York, pp 348–372
Fung A (2005) Deliberation before the revolution toward an ethics of deliberative democracy in an unjust world. Polit Theory 33(3):397–419
Gardiner SM (2011) A perfect moral storm: the ethical tragedy of climate change. Oxford University Press, Oxford
Gosseries A (2008) On future generations future rights. J Polit Philos 16(4):446–474
Greenwood M (2007) Stakeholder engagement: beyond the myth of corporate responsibility. J Bus Ethics 74:315–327
Grønbæk K, Kyng M, Mogensen P (1997) Toward a cooperative experimental system development approach. In: Kyng M, Mathiassen L (eds) Computers and design in context. MIT Press, Cambridge, MA, pp 201–238
Habermas J (1971) Knowledge and human interests. Beacon Press, Boston
Habermas J (1984) The theory of communicative action vol. I: reason and the rationalization of society (trans: McCarthy T). Beacon, Boston (German, 1981, vol 1)
Habermas J (1987) The theory of communicative action vol. II: lifeworld and system (trans: McCarthy T). Beacon, Boston (German, 1981, vol 2)
Habermas J (1990) Moral consciousness and communicative action. MIT Press, Cambridge, MA
Habermas J (1993) Justification and application: remarks on discourse ethics. MIT Press, Cambridge, MA
Habermas J (1994) Three normative models of democracy. Constellations. Int J Crit Democr Theory 1(1):1–10
Heidegger M (1977) The question concerning technology, and other essays. Harper & Row, New York (orig. 1953)
Hoffman SM, High-Pippert A (2005) Community energy: a social architecture for an alternative energy future. Bull Sci Technol Soc 25(2):387–401
Hosanagar K, Fleder D, Lee D, Buja A (2014) Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. Manag Sci 60(4): 805–823
Hösle V (1994) Philosophie der ökologischen Krise: Moskauer Vorträge. CH Beck, München
Hösle V (2004) Morals and politics. University of Notre Dame Press, Notre Dame
Howard PN, Duffy A, Freelon D, Hussain M, Mari W, Mazaid M (2011) Opening closed regimes: what was the role of social media during the Arab Spring? Project on Information Technology & Political Islam, Seattle, working paper, 2011.1
Introna LD, Nissenbaum H (2000) Shaping the web: why the politics of search engines matters. Inf Soc Int J 16(3):169–185

Introna LD (2005) Disclosive ethics and information technology: disclosing facial recognition systems. Ethics Inf Technol 7:75–86.
Johnson DG (1997) Is the global information infrastructure a democratic technology? In: Spinello RA, Tavani HT (eds) Readings in cyberethics, 2nd edn. Jones and Bartlett Publishers, Sudbury, pp 121–133
Jonas H (1979/1984) Das Prinzip Verantwortung. Suhrkamp, Frankfurt am Main/The imperative of responsibility. In search of an ethics for the technological age. The University of Chicago Press, Chicago
Kensing F, Blomberg J (1998) Participatory design: issues and concerns. Comput Supported Coop Work 7:167–185
Kensing F, Simonsen J, Bodker K (1998) MUST: a method for participatory design. Hum Comput Interact 13(2):167–198
Khondker HH (2011) Role of the new media in the Arab Spring. Globalizations 8(5):675–679
Kiran AH (2012) Does responsible innovation presuppose design instrumentalism? Examining the case of telecare at home in the Netherlands. Technol Soc 34(3):216–226
Kowalski E (2002) Technology assessment : Suche nach Handlungsoptionen in der technischen Zivilisation. vdf Hochschulverlag AG, Zürich
Kurzweil R (2005) The singularity is near: when humans transcend biology. Viking, New York
Latour B (1979) The social construction of scientific facts. Sage, Beverly Hills
Latour B (1992) Where are the missing masses? The sociology of a few mundane artifacts. In: Bijker WE, Law J (eds) Shaping technology/building society: studies in sociotechnical change. MIT Press, Cambridge, MA, pp 225–258
Latour B (2005) Reassembling the social: an introduction to actor-network-theory. Oxford University Press, Oxford/New York
Law J (1999) Actor network theory and after. Blackwell/Sociological Review, Oxford/Malden
Levidow L (1998) Democratizing technology-or technologizing democracy? Regulating agricultural biotechnology in Europe. Technol Soc 20(2):211–226
Lewis JI, Wiser RH (2007) Fostering a renewable energy technology industry: an international comparison of wind industry policy support mechanisms. Energ Policy 35(3): 1844–1857
Lim M (2012) Clicks, cabs, and coffee houses: social media and oppositional movements in Egypt, 2004–2011. J Commun 62:231–248
Manders-Huits N (2011) What values in design? The challenge of incorporating moral values into design. Sci Eng Ethics 17:271–287
Massa P, Avesani P (2007) Trust metrics on controversial users: balancing between tyranny of the majority and echo chambers. Int J Semant Web Inf Syst 3(1):39–64
Marx K (1938) Capital. London: Allen & Unwin
Mill JS (1859) On liberty. In: Robson JM (ed) Collected works of John Stuart Mill, vol 18. University of Toronto Press, Toronto, pp 213–310, 1963ff
Mill JS (1861) Considerations on representative government. Prometheus Books, Buffalo, p , 1991
Mitcham C (1990) Three ways of being-with-technology. In: Ormiston GL (ed) From artifact to habitat: studies in the critical engagement of technology. Bethlehem, PA, Lehigh University Press, pp 31–59
Mitcham C (1994) Thinking through technology: the path between engineering and philosophy. University of Chicago Press, Chicago
Morozov E (2011) The net delusion: the dark side of Internet freedom. PublicAffairs, New York
Nagel T (2005) The problem of global justice. Philos Public Aff 33:113–147
Nagenborg M (2005) Search engines, special issue of *International Review of Information Ethics*, vol 3
Nieusma D, Riley D (2010) Designs on development: engineering, globalization, and social justice. Eng Stud 2(1):29–59
Noble DF (1984) Forces of production. A social history of industrial automation. Knopf, New York

Noveck BS (2009) Wiki government: how technology can make government better, democracy stronger and citizens more powerful. Brookings Institution Press, Washington, DC
Oosterlaken I (2009) Design for development: a capability approach. Des Issues 25(4):91–102
Ornetzeder M, Rohracher H (2006) User-led innovations and participation processes: lessons from sustainable energy technologies. Energ Policy 34(2):138–150
Palm E, Hansson SO (2006) The case for ethical technology assessment (eTA). Technol Forecast Soc Change 73:543–558
Pariser E (2011) The filter bubble: what the Internet is hiding from you. Penguin Press, New York
Pogge TW (2002) World poverty and human rights: cosmopolitan responsibilities and reforms. Polity Press, London
Rawls J (1971) A theory of justice. Harvard University Press, Harvard, 1999 revised edition
Rawls J (1999) The law of peoples. Harvard University Press, Cambridge
Robertson T, Wagner I (2013) Ethics: engagement, representation and politics-in-action. In: Simonsen J, Robertson T (eds) Routledge international handbook of participatory design. Routledge, New York, pp 64–85
Roth SK (1994) The unconsidered ballot: how design effects voting behaviour. Visible Lang 28(1):48–67
Rousseau J-J (1750) The social contract and discourses. Everyman, London (1993)
Savulescu J, Bostrom N (2009) Human enhancement. Oxford University Press, Oxford
Scanlon T (1998) What we owe to each other. Harvard University Press, Cambridge, MA
Schneier B (2013a) The US government has betrayed the internet. We need to take it back. The Guardian, 5 Sept 2013. http://www.theguardian.com/commentisfree/2013/sep/05/government-betrayed-internet-nsa-spying. Accessed 2 Apr 2014
Schneier B (2013b) Why the NSA's attacks on the internet must be made public. The Guardian, 4 Oct 2013. http://www.theguardian.com/commentisfree/2013/oct/04/nsa-attacks-internet-bruce-schneier. Accessed 2 Apr 2013
Schot J, Rip A (1997) The past and future of constructive technology assessment. Technol Forecast Soc Change 54:251–268
Schudson M (1998) The good citizen. A history of American civil life. Harvard University Press, Cambridge, MA
Sclove RE (1995) Democracy & technology. The Guilford Press, New York
Sen A (1999) Development as freedom. Knopf, New York
Shelley C (2012) Fairness in technological design. Sci Eng Ethics 18:663–680
Singer P (1973) Democracy and disobedience. Oxford University Press, Oxford
Snow C (1959) The two cultures and the scientific revolution, The Rede lecture, 1959. Cambridge University Press, New York
Spahn A (2010) Technology. In: Birx H (ed) 21st century anthropology: a reference handbook. SAGE Publications, Thousand Oaks, pp 132–144
Steen M (2011) Upon opening the black box of participatory design and finding it filled with ethics. In: Proceedings of the Nordic design research conference no 4: Nordes 2011: making design matter. Helsinki, 29–31 May
Stilgoe J, Owen R, Macnaghten P (2013) Developing a framework for responsible innovation. Res Policy 42:1568–1580
Sunstein C (2001) Republic.com. Princeton University Press, Princeton
Sutherland S (1992/2007) Irrationality: the enemy within. Constable and Company/Irrationality. Pinter and Martin, London
Swierstra T, Rip A (2007) Nano-ethics as NEST-ethics: patterns of moral argumentation about new and emerging science and technology. Nanoethics 1:3–20
Taebi B, Correljé A, Cuppen E, Dignum M, Pesch U (2014) Responsible innovation as an endorsement of public values: the need for interdisciplinary research. J Responsib Innov 1(1):118–124

Tavani H (2014) Search engines and ethics. In: Zalta EN (ed) The Stanford encyclopedia of philosophy, Spring 2014 edition. http://plato.stanford.edu/archives/spr2014/entries/ethics-search/. Accessed 15 July 2014

Thompson DF (2010) Representing future generations: political presentism and democratic trusteeship. Crit Rev Int Polit Philos 13(1):17–37

Van den Hove S (2006) Between consensus and compromise: acknowledging the negotiation dimension in participatory approaches. Land Use Policy 23(1):10–17

Van den Hoven J (2005) E-democracy, E-contestation and the monitorial citizen. Ethics Inf Technol 7:51–59

Van der Velden M (2009) Design for a common world: on ethical agency and cognitive justice. Ethics Inf Technol 11:37–47

Vico G (1709) De nostri temporis studiorum ratione Lateinisch-Deutsche Ausg. Wiss. Buchges., Darmstadt (1974)

Winner L (1980) Do artifacts have politics? Daedalus 109:121–123. Also in 1986. The whale and the reactor: a search for limits in an age of high technology, 19–39. University of Chicago Press, Chicago

Winner L (1992) Introduction. In: Winner L (ed) Democracy in a technological society. Kluwer, Dordrecht, pp 1–14

Yetim F (2011) Bringing discourse ethics to value sensitive design: pathways to toward a deliberative future. AIS Trans Hum Comput Interact 3(2):133–155

Young IM (1990) Justice and the politics of difference. Princeton University Press, Princeton

Young IM (2006) Responsibility and global labor justice. J Polit Philos 12(4):365–388

# Design for the Value of Human Well-Being

Philip Brey

## Contents



**Abstract**

This chapter studies how and to what extent it is possible to design for well-being. Well-being is rarely considered in the design literature and is rarely linked to technology and design in philosophy and the social sciences. A few approaches to design for well-being have recently materialized, however, including Emotional Design, capability approaches, positive psychology approaches, and Life-Based Design. In this chapter, the notion of well-being will first be clarified and contemporary theories of and approaches to well-being will be reviewed. Next, theoretical and methodological issues in design for well-being will be discussed that must be accounted for in any successful approach. This will be followed by a review of the abovementioned four approaches to design for well-being. The chapter will conclude by considering open issues and future work in the development of design approaches for well-being.

P. Brey (✉)
Universiteit Twente, Enschede, The Netherlands
e-mail: p.a.e.brey@utwente.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_14

**Keywords**
Well-being • Design • Happiness • Emotional design • Positive psychology

## Introduction

Well-being, or quality of life, is often a central value in the design of technological artifacts, especially in the design of consumer products. Firms and designers often pride themselves with developing products that are claimed to enhance well-being, quality of life, the good life, or some similar notion. Given the centrality of well-being in much of design, one would expect an extensive literature on design for well-being. This turns out not to be the case. Very few studies in the design literature focus on well-being and even less present a methodology or approach for designing for well-being. The handful of approaches that has been developed, all of them quite recently, will be discussed in this paper later on.

In the philosophical literature, the situation is not much better. Philosophical studies of well-being very rarely mention technology or design. Conversely, the philosophy of technology only rarely focuses on the relation between technology and well-being. The philosophical studies that do focus on this relation rarely mention design. As a consequence, there is almost no connection between the few studies in the design literature on design and well-being and the few works in the philosophy literature on technology and well-being.

In the philosophy of technology literature, well-being has always been an implicit topic, rarely addressed or studied explicitly, but often presupposed as an implicit value or ideal for technology and design. Already in the Enlightenment literature on science and its application, authors like Descartes, Bacon, and Leibniz relate technology to well-being, mostly in positive terms. For example, seventeenth-century philosopher René Descartes held that the technological application of science would yield an unlimited number of devices that would allow people to effortlessly enjoy all benefits that the Earth could offer them (Descartes 1994 [1637]). The optimistic vision of technology of the Enlightenment is still present in contemporary society, in which technology is frequently conceived of as an instrument of social and economic progress that makes people's lives better.

Twentieth-century philosophy of technology, in contrast, focused on negative implications of technology for well-being. It portrayed technology as destructive of the environment and of humanity, in reference to the atrocities of Auschwitz and Hiroshima. It thematized rationalization, alienation, materialism, and loss of community as implications of a technological society. It argued that technology had gone out of control and that humanity was made subservient by it. These were themes in the early and mid-twentieth century, put forward by authors like Theodor Adorno, Max Horkheimer, Herbert Marcuse, and Martin Heidegger, as well as in the mid- to late twentieth century by authors like Jacques Ellul, Albert Borgmann, Hubert Dreyfus, Jean-François Lyotard, and Jean Baudrillard. Some of these critical philosophers did raise the possibility of transforming technology so as to

be more supportive of well-being and other human values, for example, Ivan Illich, Langdon Winner, and Andrew Feenberg. It is only in the early twenty-first century, however, that philosophers have started to explicitly study the relation between well-being, technology, and design (Van de Poel 2012; Brey et al. 2012).

The relation between technology, well-being, and design is also rarely studied in the social and behavioral sciences. In recent decades, there has been a great deal of interest in well-being in the social sciences, especially in psychology and economics, but the discussion rarely turns to the role of technology. It sometimes does so indirectly, because of a recurring focus on consumer culture, which revolves around technologically produced consumer products. Social scientists have developed mostly critical assessments of consumer culture, its products, and its implications for well-being (Lebergott 1993; Kasser and Kanner 2004; Dittmar 2008). Recently, social scientists have also begun to study the implications of information technologies for well-being (Amichai-Hamburger 2009; Turkle 2011). Rarely do these studies consider the topic of design, however.

In the following sections, I will study the very idea of design for well-being. I will begin, in the next section, by clarifying the notion of well-being and by considering theories of it and approaches to its study. I will then discuss theoretical and methodological issues in designing for well-being that must be accounted for in any successful approach. Next, I will review and critique four current approaches to design for well-being. I will conclude by considering open issues and future work in the development of design approaches for well-being.

## What Is Well-Being?

Well-being is a state of persons which designates that they are happy or flourishing and that their life is going well for them. Well-being is often considered to be the highest value to which other values can be subsumed: it is that what makes one's life good; a good life is for many people of the highest value. In some of the most important systems of ethics, most notably utilitarianism, well-being (or "utility") is considered to be the highest good, and acts are to be morally evaluated according to the aggregate amount of well-being that they bring about. Well-being is sometimes equated with happiness, but not all theorists agree. Happiness, on most accounts, consists of the presence of positive feelings (and the absence of negative ones). As we will see, some philosophers have argued that well-being implies more than just having positive mental states.

Well-being has been studied by philosophers since the ancient Greeks. The philosophical study of well-being over many centuries has yielded three major types of theories of well-being: hedonist, desire-fulfillment, and objective list theories (Parfit 1984; Brey 2012). I will now briefly discuss them in turn.

*Hedonist theories* hold that a well-being consists of the presence of pleasure and the absence of pain. A good life is a life in which one successfully accumulates pleasurable feelings and avoids painful ones. Although hedonist philosophies can be traced back to the ancient Greeks, contemporary hedonism finds its roots in

eighteenth and nineteenth century utilitarianism. A distinction is often made between two types of hedonism. The first of these, *quantitative hedonism*, was originally proposed by Jeremy Bentham (1996 [1789]). It holds that the value of pleasure for one's life is only determined by its quantity, as measured among others by the duration and intensity of the feeling, and not by its quality.

Quantitative hedonism has been criticized for putting satisfaction of primitive urges at the same footing as more sophisticated pleasures, such as those resulting from friendship and art. The life of a pig is just as good as the life of a refined person, on this conception, as long as the amount of pleasure for both is the same. John Stuart Mill (1998 [1863]) argued against this position that certain types of pleasure are more desirable or worth having than others. This applies to the so-called higher pleasures, such as experiences of friendship, knowledge, art, contemplation, and refinement in taste. His *qualitative hedonism* holds that a good life is not just a life with many pleasant experiences but a life with many experiences of the so-called higher pleasures.

A problem for hedonist theories is that there seem to be qualities that make our life better which cannot be reduced to pleasure. These include the authenticity or veracity of our experiences. A life which is pleasant but which is built on illusions and deceptions would seem to be worse than a somewhat less pleasant life which is fully authentic. This problem of authenticity does not occur with *desire-fulfillment theories*, also called *preference-satisfaction theories*. Such theories hold that well-being lies in the fulfillment of one's desires. Desire-fulfillment theories emerged in the nineteenth century, in part as an outgrowth of welfare economics. Several versions have been proposed, from basic to more sophisticated (Crisp 2013).

The simplest version, so-called *simple desire-satisfactionism*, holds that people are better off to the extent that their current desires are fulfilled. A problem with this view is that many desires of people seem to go against their long-term interests. People are often mistaken about what is good for them in the long run, and they often act on impulses that they later regret. *Informed desire-fulfillment theories* overcome this problem by claiming that the best life one could lead is the life in which those desires are fulfilled that one would have if one were fully informed of one's situation. If one is properly informed, one would only desire those things that make a good fit with one's character and that one is likely to be able to realize.

*Objective list theories* of well-being hold that well-being is the result of a number of objective conditions of persons and do not rest on a person's subjective preferences or experiences. There are objective goods that contribute to a person's well-being even if that person does not desire them or experiences pleasure from them. Well-being is attained when one has attained most or all goods on the list. Goods that are often put forward in objective list accounts include liberty, friendship, autonomy, accomplishment, wisdom, understanding, morality, the development of one's abilities, enjoyment, and aesthetic experience (Parfit 1984; Griffin 1986). One influential type of objective list account, *perfectionism*, holds that what is good for us is given by our intrinsic nature as human beings and that we should strive to excel so as to realize these goods. The most famous perfectionist theory is Aristotle's theory of *eudaimonia* or flourishing.

The second half of the twentieth century has seen the emergence of theories of well-being in the fields of economics and psychology (Brey 2012). Most of these theories can be placed within the categories that have been distinguished in philosophy. In psychology, a true *psychology of happiness* started to emerge in the 1980s, due in part to the seminal work of Ed Diener, who strived to make well-being a measurable quantity that could be studied empirically (Diener 1984). Since Diener, psychologists tend to focus on *subjective well-being*, which is how people evaluate their own happiness, something that can be recorded and measured. They also use the term *life satisfaction*, which denotes how people assess the balance between positive and negative affect and success and failure in their lives as a whole and which is to be contrasted with assessments of subjective well-being which denote people's happiness at a certain point in time (Diener and Lucas 1999). The 1990s subsequently saw the emergence of *positive psychology* (Seligman and Csikszentmihalyi 2000), an approach within the psychology of happiness that has become dominant. Positive psychology does not merely aim to study well-being but also to develop psychological techniques and tools for making people's lives more fulfilling.

In economics, happiness and well-being have been important topics since nineteenth-century neoclassical economics, which explains economic activity in terms of its expected utility (which is often equated with well-being), and twentieth-century welfare economics, which aims to measure social welfare and improve it through economic solutions that maximize utility. An important approach within contemporary welfare economics, which has been taken up in philosophy as well, is the *capability approach* (Sen 1980). This approach assumes that people's ability to attain well-being depends on their possession of a number of basic capabilities, the development of which can be supported through social and economic means. It will be discussed more extensively in section "Four Design Approaches." Recent decades have seen the emergence of *happiness economics*, a new branch of economics that studies the economic conditions for happiness and well-being and that relies strongly on psychological research on happiness (Bruni and Porta 2005).

## What Does It Mean to Design for Well-Being?

In this section, several theoretical and methodological issues will be discussed regarding design for well-being. The first issue is whether technological artifacts are capable of promoting or enhancing well-being and whether it is possible to design for well-being. Having answered these two questions positively, we will then move on to the question of how different conceptions of well-being, as discussed in section "What is Well-Being?," can be related to technological designs. Next, we will discuss the epistemological problem of how designers are to know what conception of well-being they should design for and how to find this out by studying users and stakeholders. We will then consider the scope problem, which concerns the question of how to delineate both the people whose well-being will be considered and the possible effects on well-being that will be considered.

The aggregation problem will be next, which concerns the question of how multiple and possibly conflicting well-being values can be accounted for in a single design. Finally, we will consider the specification problem, which is how to derive specifications for designs that promote well-being.

The first issue we will consider is whether design for well-being is actually possible. Several approaches in ethics of technology hold that this is. They hold that the use of technological artifacts can often be reliably correlated with consequences or effects beyond the intended function of the artifact. Some of these consequences will be desirable, whereas others will not be. It may be possible to design artifacts in such a way that a certain desired consequence is bolstered or an undesirable consequence is avoided. Increased well-being is a possible consequence of the use of a technological artifact. Therefore, it is possible, in principle, to design for well-being.

The most notable approach in ethics of technology that holds this position is the approach of value-sensitive design (Friedman et al. 2006; Brey 2010; cross-reference to chapter "► Value Sensitive Design: Applications, Adaptations, and Critiques") or VSD. The VSD approach explicitly takes values into account in design and aims to arrive at designs that adhere to, or promote, these values. Well-being is recognizable as a value: it is an abstract ideal that is part of our conception of the good. In VSD approaches, well-being is recognized as an important value that can be incorporated into designs. Within the VSD approach, to say that an artifact embodies a value such as well-being is not to say that the artifact will deterministically bring about well-being, however and by whomever it is used. Rather, it is to say that given a particular user or range of users, and in a particular context or range of contexts, the use of the artifact will tend to promote well-being. Given one's knowledge of user and context of use, it is therefore possible to design for well-being.

Design for well-being means different things for different conceptions of well-being, as discussed in section "What is Well-Being?." On a hedonist conception, design for well-being is design for pleasurable experiences or for the prevention or lessening of negative ones. By this criterion, the focus will be on designing artifacts that cause pleasant sensations, enable users to undertake activities that are pleasurable, allow users to avoid unpleasant activities, and prevent or reduce mental and physical pain and discomfort. On a desire-fulfillment conception, design focuses on satisfying desires of users and other stakeholders. Within a simple desire-fulfillment approach, this would require investigations into what people desire and then creating designs to help these desires come true (and to avoids side effects that are not found desirable). Design is more difficult within an informed desire-fulfillment approach, since it requires determination of the hypothetical desires that people would have if they were properly informed, which is a more speculative endeavor. On an objective list conception, finally, designs should be such that they help bring about the acquisition of goods on the objective list. For example, they should support the development and maintenance or exercise of friendships, autonomy, and practical wisdom.

Having claimed that design for well-being is possible, I will now turn to several problems that a design approach for well-being must overcome.

The first of these is the *epistemological problem* (Van de Poel 2012). This problem is how designers are to know which conception of well-being applies to a particular user or stakeholder. On most conceptions of well-being, what is well-being for a particular person cannot be determined objectively, through objective criteria independently from that person. It requires an understanding of that person, which may include knowledge of his or her preferences, desires, values, traits, and social and cultural embeddedness. In the absence of this knowledge, designers need to know what constitutes well-being for the expected user(s) of the design. Design for well-being may therefore require studies of users, such as questionnaires, tests, or experiments, to reveal their preferences, values, or other traits. But how is this to be done?

Another problem in designing for well-being is what I call the *scope problem*: what range of stakeholders and potential consequences for well-being will be considered in the design? One issue here is the extent to which indirect effects on well-being are considered. A video game may have immediate positive effects on well-being, such as feelings of joy and excitement. But it may also have indirect negative effects, such as social isolation, sleep disturbances, and strain to arms and neck. It would seem that consideration of such indirect effects should be included in the equation in design for well-being. However, indirect effects are often more difficult to determine and more contingent on other factors. It therefore needs to be considered to which extent indirect effects will be considered.

Another scope issue is whether design for well-being should focus on users (user well-being) or also on other stakeholders (stakeholder well-being). From an ethical point of view, taking into account the well-being of all stakeholders seems preferable over just accounting for user well-being, although Van de Poel (2012) argues that designers normally do not have a moral imperative to account for stakeholder well-being. A final scope issue concerns the definition of well-being in terms of its duration. In section "What is Well-Being?," a distinction was made between subjective well-being and life satisfaction. The effect of using an artifact on well-being may be measured by focusing on relatively immediate positive and negative experiences of users. Another measure is to consider how using the artifact over a longer period of time may change people's assessments of the quality of their lives as a whole. This will be more difficult to measure, but will ultimately be more important than immediate effects on subjective well-being.

Next, the *aggregation problem* is described by Van de Poel (2012) as the problem of how the well-being of different people can be aggregated into an overall measure of well-being. The aggregation problem exists because different people will often have different and possibly even conflicting or incommensurable conceptions of well-being, and this raises the question of how different well-being values of different stakeholders can be added up and combined into one measure of well-being and one design. This problem applies to cases of design for the well-being of stakeholders beyond the user and also to cases in which artifacts may be used by multiple users with different well-being values. Not recognized by Van de Poel, it may also emerge in accounting for the well-being values of a single user, which may be conflicting and incommensurable as well.

Finally, there is the *specification problem*, which is how to go from well-being values to design specifications. Such design specifications define design features (structural features, capabilities, affordances) that tend to correlate with positive effects on well-being when the artifact is used by a particular class of users in a particular class of contexts. These design features have a causal role, together with other causes in the context of use, in bringing about states of well-being. For example, textural smoothness in a hand-held device may bring about pleasant feelings that may be absent in a device without this feature. Although there are probably rules of thumb for relating well-being values to design features, determining such features is likely to be part of the creative process in design.

## Four Design Approaches

Design for well-being has only quite recently become a subject that designers, philosophers, and social scientists have started paying attention to. Little work yet exists in this new area, and the approaches that have been developed are still young. In what follows, I will present and discuss four distinct approaches that have recently been put forward.

### Emotional Design

*Emotional Design* is the name for a family of approaches that use design to evoke emotional experiences in users. It focuses on emotional experiences of users with products and emotional meanings associated with product use. The aim of Emotional Design is to provide products with additional utility by designing them to evoke pleasure and other positive emotions in users. The term "Emotional Design" was coined by Donald Norman in 2005.

Patrick Jordan was one of the first designers to focus on positive emotions in design in an approach that he called *pleasure design*. Jordan (2000) claims that designers should not just design for functionality and usability but also for pleasure: products should be pleasurable to use. He distinguishes four types of pleasure which he claims motivate humans to use products. Although he claims that pleasurability is not a property of a product but the result of the interaction of users with a product, he claims that designers could design affordances in product that would make them pleasurable to use for most users. He therefore associates the four pleasures with distinct design features.

Jordan distinguishes the following four pleasures and associated design features:

- *Physio-pleasure*: Bodily pleasure deriving from the sense organs (touch, taste, smell, appearance, and sound). Associated design features: pleasurable sensory features.
- *Psycho-pleasure*: Pleasure deriving from cognitive and emotional reactions. Associated design features: usability.

- *Socio-pleasure*: Pleasure arising from one's relationship with other people or society as a whole. Associated design features: markers of social or cultural status, features that express social messages.
- *Ideo-pleasure*: Pleasure arising from people's values and tastes – cultural and aesthetic values, moral values, and personal aspirations. Associated design features: material or semiotic features that express values like sustainability, sobriety, artistic values, religious values, etc.

According to Jordan, then, designers can enhance the well-being of users by equipping products with design features that enable pleasurable feelings of these four kinds.

In his 2005 book *Emotional Design*, design guru Donald Norman introduces Emotional Design as an approach that takes into account the emotional response of users to products and that strives for products to evoke positive feelings. Norman's focus is on positive emotions, which resemble Jordan's pleasures, but possibly denote a broader range of feelings. Norman claims that users experience emotions in products at three levels, each of which can be accommodated for by different design features:

- *Visceral level*: At this level, people have rapid responses to products, making rapid judgments regarding goodness, badness, safety, and danger. Visceral responses are biologically determined and involve signals to the muscles and to the rest of the brain. They constitute the start of affective processing. Associated design features: general appearance.
- *Behavioral level*: This is the level at which people experience the use of products. At this level, people are not concerned with appearance, but with the usability and effectiveness of the product. Relevant components at the behavioral level are function, understandability, usability, and physical feel. Associated design features: functional properties, usability, and tactile features.
- *Reflective level*: This level is the least immediate. It is the level at which the product is attributed a meaning beyond its appearance and use. At this level, products and their use evoke meaning, culture, self-image, personal remembrances, and messages to others. Associated design features: material and semiotic features that express such meanings.

In Norman's approach, therefore, well-being is enhanced by designing products so that they evoke positive emotions associated with appearance, use and their broader personal, social, and cultural significance.

A third author that deserves mention in the context of Emotional Design approaches is psychologist Mihaly Csikszentmihalyi, whose 1990 book *Flow* has had a major influence on product design. Csikszentmihalyi claims that when people engage in activity, they may reach a state of flow, which is an emotional state between boredom and anxiety. Flow is a feeling of complete and energized focus in an activity, with a high level of enjoyment and fulfillment. In flow, there is a good balance between challenge and skill, so that the task at hand is neither too

challenging nor not challenging enough, and someone's skills are experienced as making a good fit with the task at hand. This provides one with a feeling of flow. Flow has been used as a design criterion for computer interfaces, websites, and computer games, among others, with the aim of designing an adequate balance between challenge and skill for intended user groups (King 2003; Chen 2007).

Emotional Design approaches constitute a strong attempt to go beyond mere functionality of products to consider positive feelings in product design. While such positive feelings may enhance well-being, it would go too far to consider these approaches as constituting a comprehensive design approach for well-being. Emotional Design focuses on positive feelings that are evoked by artifacts while they are used or perceived. There is little or no attention to consequences for well-being beyond these immediate experiences. These include indirect pleasurable feelings that indirectly result from the use or possession of a product and more lasting effects on life satisfaction. Therefore, Emotional Design approaches are at best only a component of a comprehensive approach to design for well-being and not a comprehensive approach in itself.

## Capability Approaches to Design

Capability approaches to design (Oosterlaken 2009; Johnstone 2012; Oosterlaken and van den Hoven 2012) take an approach that is very different from Emotional Design. Rather than focusing on the arousal of positive feelings, they focus on the enhancement of people's basic capabilities for leading a good life. The foundation for this approach is found in the capability approach, an approach to well-being and welfare that rests on the assumption that people's ability to attain well-being is dependent on their development and possession of a number of basic capabilities that allow them to engage in activities that promote their well-being. The capability approach has originally been developed by economist Amartya Sen (1980) as an approach to welfare economics and has been further developed in philosophy by Martha Nussbaum, one of Sen's collaborators (Nussbaum 2000).

The capability approach assumes that well-being is possibly the most important condition for people to strive for and that every person should have the freedom to achieve well-being. It then goes on to claim that in order to attain a state of well-being, people must be in possession of a set of basic capabilities, which are opportunities to do things or be things that are of value to them. The exact set of capabilities may vary from person to person, since people may have different conceptions of value that require different sets of capabilities for their realization. However, Martha Nussbaum (2000) has argued that people tend to be in agreement on a basic set of ten capabilities. These include capabilities to life, bodily health, emotions, practical reason, play, affiliation, control over one's environment, and others. Being in possession of a basic set of capabilities is no guarantee for well-being; these capabilities must also be exercised in order to realize the actions or states of being that the capabilities are directed at. But they are necessary conditions for well-being, since people cannot attain well-being without them.

The relevance of this approach to design is as follows. Technological artifacts can strengthen and extend human capabilities. For example, technological products can help people find, prepare or consume foods, help them exercise, or help them protect themselves from or heal from injuries or disease, thereby helping them stay alive and maintain bodily health. Or, in relation to Nussbaum's emphasis on play as a component of well-being, technological products can help people play and enjoy recreational activities. Design for well-being, according to the capability approach, is therefore designing products that enhance basic capabilities for well-being.

Capability approaches prescribe that products should be designed to enhance one or more basic capabilities and to avoid harm to the other capabilities. They will tend to emphasize that the functionality of an artifact should be directed the enhancement of basic capabilities for a good life, rather than to other functionalities that cannot clearly be related to the capabilities needed for leading a good life. The enhancement of basic capabilities can either be found in the proper function of an artifact or in secondary functionalities or features.

Capability approaches have the advantage over Emotional Design approaches that they can better account for indirect and lasting effects of product use for well-being. In fact, they have little concern for the fleeting feelings of pleasure that Emotional Design approaches focus on. Instead, they focus on building and expanding durable capabilities that are core requisites for lasting well-being.

In spite of this advantage of capability approaches to design, there are also significant weaknesses to them. First, there is considerable disagreement among proponents of the capability approach about whether it is possible to draw up a list of basic capabilities that applies to most or all people, and if so, what this list should look like. Even if there were to be agreement on a particular list, the capabilities may be so abstract that they do not offer sufficient guidance for design. One reason is that it may be unclear whether a design makes a net positive contribution to a capability. Do mobile phones, for example, contribute to better affiliations with others? It depends on what one takes as the default situation against which they are measured. Another reason is that it may be unclear whether a positive contribution of a design to a capability outweighs a negative contribution to a different one. For example, a motorized bicycle (as compared to an unmotorized one) may give one more control over one's environment by increasing one's mobility, but may also negatively affect one's physical health by reducing one's physical activity level. How is one to determine whether the design makes a net positive contribution to well-being?

Overall, capability approaches to design seem promising, but the guidance to actual design practices has so far been limited because of the abstract, unoperationalized nature of existing approaches. Perhaps there is also a more fundamental issue here: capabilities are decontextualized phenomena, and because of this, capability approaches may have difficulties taking into account the context in which capabilities are used and the particular values and characteristics of the persons who have them. Perhaps this issue can be mitigated through the further development of Sen's (1992) notion of "conversion factors," which are contextual factors that help determine the way in which resources (including technologies)

give shape to capabilities. A different limitation of the capability approaches is that they seem to ignore consequences of the use of artifacts for well-being that are not mediated by capabilities. For example, using an artifact may evoke feelings of pleasure or contentment that do not involve the augmentation of a particular capability but that nevertheless contribute to well-being. Clearly, it is not only capability-enhancing features of an artifact that affect well-being; all kinds of other impacts of using an artifact may affect it as well.

## Positive Psychology Approaches

Positive psychology is an approach within psychology that focuses on studying and improving people's positive functioning and well-being (Seligman and Csikszentmihalyi 2000). Positive psychology has recently emerged as the dominant psychological approach to well-being. It stands apart from many other psychological approaches through its focus on the enhancement of creativity, talent, and fulfillment, rather than on treatment of mental illness.

A very influential theory in positive psychology is Martin Seligman's theory of authentic happiness (Seligman 2002). Seligman holds that a good life is a combination of three types of lives: the pleasant life, the engaged life, and the meaningful life. The pleasant life is attained by having, and learning to have, positive feelings, that are directed at the present, past, and future. The engaged life consists in the pursuit of engagement and involvement in work, intimate relations, and leisure. Engaged activity brings experiences of flow (Csikszentmihalyi 1990) in which one's attention is completely focused and time loses meaning. The pursuit of engaging activities requires character strengths and virtues ("signature strengths") that allow one to execute activities in an engaged manner. The meaningful life, finally, is a life in which one's signature strengths and talents are used in the service of things that one believes to be bigger than oneself.

Although not much work has yet been done to apply positive psychology to design, an interesting approach has been developed by Ruitenberg and Desmet (2012). They want to use positive psychology to design products and services that promote happiness or well-being. They focus on long-term life satisfaction rather than short-term experiences or emotions in using products. As they explain, they want to make a shift from product experience (the focus of Emotional Design approaches) to *meaningful activities*. Meaningful activities are activities that use and develop personal skills and talents of the users, that are rooted in core values of the user, that contribute to a greater good (a thing or person), and that are rewarding and enjoyable in themselves.

Ruitenberg and Desmet want to design products that enable and inspire people to engage in meaningful activities, as defined above, that contribute to happiness and life satisfaction. Design includes visualizing meaningful activities and then designing products that enable or inspire people to engage in these activities. They recognize product-oriented design strategies, as found in Emotional Design approaches, as part of their approach. As they state, pleasurable experiences

associated with product use can contribute to overall happiness. However their major focus is on meaningful activities that are enabled or suggested by products and that either contribute to happiness or take away sources of unhappiness. As an example of a product that stimulates meaningful activities, they describe a "vegetable book," a large book placed on a stand in a communal vegetable garden with a page for each vegetable which allows users to place notes on how to grow or cook the vegetable. This product enables and stimulates sharing activities between gardeners that contribute to overall happiness.

Their approach is aimed at inducing behavioral change by stimulating voluntary changes in our daily actions. Our daily routines should become tied to proven strategies for increasing happiness such as "cultivating optimism," "nurturing relationships," "taking care of your body," and "practicing acts of kindness." Products should be designed to support behavioral changes toward such strategies. This is not easy and requires that people are already receptive and motivated to change their routines, know how to do it, and are triggered in some way to do it. Ruitenberg and Desmet believe that good designs can provide the necessary triggers and can support people in being motivated and enabled to make behavioral changes.[1]

The positive psychology approach of Ruitenberg and Desmet has as strong points that it focuses on long-term life satisfaction while also incorporating the temporary positive experiences sought in Emotional Design and that they aim to develop explicit design strategies for triggering behavioral changes in persons toward more wholesome routines. The approach is still in its early stages, however. The design methodology and the conceptual framework for identifying dimensions of well-being that can be incorporated into meaningful activities are still underdeveloped. Perhaps these limitations can be overcome in the future. A potentially more fundamental criticism is that this approach focuses on relatively isolated behavioral routines and does not consider the whole lives in which these routines are supposed to function and to contribute to an overall ideal of well-being. The next design approach that will be discussed does have whole lives as its focus.

## Life-Based Design

Life-Based Design (Leikas 2009; Saariluoma and Leikas 2010; Leikas et al. 2013) is a design approach that aims to improve well-being by looking at people's whole lives and the role of technologies in them. Looking at whole lives involves studying people's forms of life, values, and circumstances and taking these into account in

[1]Desmet and Pohlmeyer (2013) develop a similar approach within positive psychology which they call *positive design*. Positive design aims to design for pleasure (personal affect), personal significance (pursuing personal goals), and virtue (being a morally good person). All three should be strived for in each design. Such designs are then held to enhance overall well-being.

design. The notion of a *form of life*, originally proposed by Ludwig Wittgenstein, is key to the approach. A form of life is a practice or "system of rule-following actions," such as a hobby, an activity, a profession, or a family role. Examples are "motor cycling," "being a soccer fan," "being a grandparent," "being a medical doctor," and "living in a senior home."

There are four phases in Life-Based Design (Leikas et al. 2013):

1. *Form of Life Analysis*. In this phase, a particular form of life is described, including the rule-following actions and practices that are typical to it, the values that people share in the form of life, and typical actors, contexts and actions, and explanations of their connectedness. Problematic issues in the form of life are also identified, and design goals are formulated, ultimately resulting in a set of human requirements which define in general terms how people's lives in a specific form of life should be improved.
2. *Concept Design and Design Requirements*. In this phase, a more precise definition of the problem to be solved is developed and it is conceived how this problem may be solved through a technological design. It states how a technology may achieve action goals that are believed to make the user's life better and ends up with a technical design and implementation. After this phase, it is clear what the technology looks like and how people will use it in their lives.
3. *Fit-For-Life Design*. In this phase, it is investigated, in interaction with users, whether the proposed design ideas do really add to their quality of life and whether the technological solution chosen is optimal. This is an iterative process that may lead to repeated improvements in the technology.
4. *Innovation Design*. In this final phase, procedures are developed and implemented for incorporating the new technology into real-life settings and making it ready for general use. This includes accounts of the social and technical infrastructure for the technology, development of a marketing plan, and further needed auxiliary activity.

Life-Based Design has several strong features compared to the other three approaches that have been discussed. It takes a more integral and more contextualized approach than Emotional Design and Capability approaches. Compared to positive psychology approaches, it has a somewhat broader focus as well, focusing on forms of life rather than the often more specific behavioral routines that are the focus of positive psychology approaches. Life-Based Design is still novel, however; its methodology needs to be developed more and few case studies have been developed to apply the theory. It currently lacks a conceptual framework for identifying well-being values that are at stake in the forms of life it studies. A potential weakness is that its focus is on improving existing forms of life and does not seem to include the possibility of developing new forms of life. It is in this way complementary to the positive psychology approach that was discussed, which aims to develop brand-new behavioral routines.

## Conclusion: Open Issues and Future Work

Design for well-being is still in its infancy. Only recently, since the year 2000, has a handful of approaches emerged, which are still limited in scope and require further methodological development and which have not yet been developed extensively using case studies. Our discussion has indicated that design for well-being is possible but must deal with complex issues that have not yet been adequately resolved, which include the scope problem, the epistemological problem, the aggregation problem, and the specification problem. The design approaches that were reviewed recognize and engage these problems to a limited extent only.

Of the four approaches that we discussed, Emotional Design is well-developed but is the most limited in scope, focusing mostly on product experience. Capability approaches are complementary to Emotional Design approaches in their focus on capability building, but it is often unclear what the relevant capabilities are for design, and the approach is rather decontextualized. The positive psychology approach has a strong point in that it stimulates new meaningful practices. Life-Based Design is the most comprehensive existing approach but is conservative in aiming to improve existing forms of life only. There seems to be a complementarity in the four approaches, in that they emphasize different factors in well-being: product experience for Emotional Design, capability building for capability approaches, stimulating new meaningful practices for the positive psychology approach, and improving existing practices for Life-Based Design. Perhaps, then, a combination of these four approaches is needed for a fully comprehensive design approach for well-being.

Any design approach should solve the epistemological problem of identification of relevant well-being values for particular users or user groups. I believe that this problem can be solved in two steps. The first step is the development of a list of well-being values for humanity as a whole. Such a list could be based on broad surveys combined with conceptual analysis and would be composed of well-being values that hold for at least some groups in society, with adequate descriptions and operationalizations. This list will possibly contain items such as autonomy, deep personal relationships, engagement, play, and achievement. A second step is to do empirical investigations of particular users or user groups so as to find out which of these well-being values apply to them (in a particular context or in relation to a particular problem). There are standard empirical protocols to do this in the field of happiness psychology. Current design approaches make little use of the extensive research approaches that have been developed in happiness psychology and related fields, and design practices are likely to improve if they start doing so. The result of this second step is an understanding of the well-being values that apply to particular users in particular contexts, which is needed to solve the epistemological problem.

Similar progress can also be made on the scope problem. Ideally, from the point of view of optimizing well-being for society as a whole, design would focus on stakeholder well-being and on life satisfaction, rather than users only and transient positive experiences only, and it would consider indirect effects on well-being as well as direct ones. For practical purposes, it may often be necessary to opt for a

more limited scope. Research is needed to determine what scope should be chosen in particular situations from a standpoint of feasibility and cost-effectiveness.

The aggregation problem may be approached in one of two ways. When it occurs because of value pluralism within users, it can be avoided by making artifacts configurable for different user groups in a way that reflects their values or by designing different versions of an artifact for different user groups. A second way to deal with the problem is to develop a framework for resolving conflict and incommensurability between different well-being values. For individual users, this may perhaps be achieved by asking them to assign weight to different well-being values and rank them in relation to each other. For groups of users, the same can be done, but in addition, procedures must be developed for resolving conflicts of interest between their members. This could be done through ethical analysis or democratic decision-making processes.

As I argued earlier, the specification problem can never be fully solved, because design is a creative process that cannot be fully captured in rules. But it may be possible to develop a set of design principles that specify, for different well-being values, which kinds of design solutions may work for them. For example, the Emotional Design approach suggests that it may be possible to list materials, shapes, colors, textures, and so on, which typically evoke positive feelings, or which do so contingent on further characteristics of the user and the context of use. Similarly, there may be design principles that specify what type of communication a communication technology must support in order to support friendships or which types of information processing do or do not support autonomy by supporting autonomous choices. Further development of methods of design for well-being may result in a large arsenal of design principles that help designers go from well-being values to design specifications.

Technology plays a powerful role in our lives and is a key factor in well-being. Well-being is one of our highest values, a value that subsumes many others. Developing approaches to design for well-being should therefore be a major goal in design. Design for well-being is feasible, but much progress is still to be made in developing approaches for it. I have reviewed current approaches and have indicated how progress can be made toward the development of sophisticated, comprehensive, and effective approaches to design for well-being.

## References

Amichai-Hamburger Y (ed) (2009) Technology and psychological well-being. Cambridge University Press, New York

Bentham J (1996 [1789]) In: Burns J, Hart HLA (eds) An introduction to the principles of morals and legislation. Clarendon, Oxford

Brey P (2010) Values in technology and disclosive computer ethics. In: Floridi L (ed) The Cambridge handbook of information and computer ethics. Cambridge University Press, Cambridge, pp 41–58

Brey P (2012) Well-being in philosophy, psychology and economics. In: Brey P, Briggle A, Spence E (eds) The good life in a technological age. Routledge, New York, pp 15–34
Brey P, Briggle A, Spence E (eds) (2012) The good life in a technological age. Routledge, New York
Bruni L, Porta P (2005) Economics and happiness: framings of analysis. Oxford University Press, Oxford
Chen J (2007) Flow in games (and everything else). Commun ACM 50:31–43
Crisp R (2013) Well-being. In: Stanford Encyclopedia of philosophy. Accessed on 2 Sep 2014 at http://plato.stanford.edu/entries/well-being/
Csikszentmihalyi M (1990) Flow: the psychology of optimal experience. Harper Perennial, New York
Descartes R (1994 [1637]) Discourse on method. Everyman Paperbacks, New York
Desmet P, Pohlmeyer A (2013) Positive design: an introduction to design for subjective well-being. Int J Design 7(3):5–19
Diener E (1984) Subjective well-being. Psychol Bull 95:542–575
Diener E, Lucas RE (1999) Personality and subjective well-being. In: Kahneman D, Diener E, Schwarz N (eds) Well-being: the foundations of hedonic psychology. Russell-Sage, New York, pp 213–229
Dittmar H (2008) Consumer society, identity, and well-being: the search for the 'Good Life' and the 'Body Perfect'. Psychology Press, London/New York
Friedman B, Kahn P, Borning A (2006) Value sensitive design and information systems. In: Zhang P, Galletta D (eds) Human-computer interaction in management information systems: foundations. M.E. Sharpe, Armonk
Griffin J (1986) Well-being: its meaning, measurement and moral importance. Clarendon, Oxford
Johnstone J (2012) Capabilities and technology. In: Brey P, Briggle A, Spence E (eds) The good life in a technological age. Routledge, New York, pp 77–91
Jordan P (2000) Designing pleasurable products. Taylor & Francis, London
Kasser T, Kanner A (eds) (2004) Psychology and consumer culture: the struggle for a good life in a materialistic world. American Psychological Association, Washington, DC
King A (2003) Speed up your site: web site optimization. New Riders Press, Indianapolis
Lebergott S (1993) Pursuing happiness: American consumers in the twentieth century. Princeton University Press, Princeton
Leikas J (2009) Life-based design – a holistic approach to designing human-technology interaction, VTT Publications 726. Edita Prima, Helsinki
Leikas J, Saariluoma P, Heinilä J, Ylikauppila M (2013) A methodological model for life-based design. Int Rev Soc Sci Human 4(2):118–136
Mill JS (1998[1863]) In: Crisp R (ed) Utilitarianism. Oxford University Press, Oxford
Norman D (2005) Emotional design: why we love (or Hate) everyday things. Basic Books, New York
Nussbaum M (2000) Women and human development: the capabilities approach. Cambridge University Press, Cambridge
Oosterlaken I (2009) Design for development: a capability approach. Design Stud 25(4):91–102
Oosterlaken I, Van den Hoven J (eds) (2012) The capability approach, technology and design. Springer, Dordrecht
Parfit D (1984) Reasons and persons. Oxford University Press, New York
Ruitenberg H, Desmet P (2012) Design thinking in positive psychology: the development of a product-service combination that stimulates happiness-enhancing activities. In: Brassett J, Hekkert P, Ludden G, Malpass M, McDonnell J (eds) Out of control; Proceedings of the 8th international conference on design and emotion, Central Saint Martins College of Art & Design, London, pp 1–10
Saariluoma P, Leikas J (2010) Life-based design – an approach to design for life. Glob J Manage Bus Res 10(5):17–23

Seligman MEP (2002) Authentic happiness: using the new positive psychology to realize your potential for lasting fulfillment. Free Press, New York
Seligman MEP, Csikszentmihalyi M (2000) Positive psychology: an introduction. Am Psychol 55:5–14
Sen A (1980) Equality of what? In: McMurrin SM (ed) The Tanner lectures on human value. University of Utah Press, Salt Lake City, pp 195–220
Sen A (1992) Inequality Re-examined. Clarendon, Oxford
Turkle S (2011) Alone together: why we expect more from technology and less from each other. Basic Books, New York
Van de Poel I (2012) Can we design for well-being? In: Brey P, Briggle A, Spence E (eds) The good life in a technological age. Routledge, New York, pp 295–306

# Design for the Value of Inclusiveness

Simeon Keates

## Contents



**Abstract**

There is an increasing awareness that many everyday products and services present challenges and difficulties to potential users. These difficulties may arise because the products and services have not been designed to allow for the full range of functional capabilities of the users who wish to use them. Medical conditions, accidents, ageing, or genetic predisposition means that most people will at some point experience functional impairments that make everyday products and services difficult to use. This chapter aims to introduce readers to the needs of the full range of users and provide an introduction to how they can develop more inclusive products and services. It addresses the principal approaches and tools to designing for inclusivity as well as the underlying rationale for why companies and designers need to consider this important set of users.

**Keywords**

Universal design • Inclusive design • User-centered design • Disability • Ageing • Impairments

S. Keates (✉)
School of Engineering, University of Greenwich, London, UK
e-mail: s_keates@yahoo.co.uk



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_15

## Introduction

What makes a successful product? This is a question that designers and design commissioners ponder regularly. In many cases, the question can be answered by reference to a property or attribute of the design, for example, the fastest, the cheapest, the most reliable, etc. Such attributes can be measured and quantified by direct empirical analysis.

However, when designing for use by consumers, i.e., people, the attributes commonly cited suddenly become decidedly more woolly and imprecise. Words and phrases such as user-friendliness, intuitiveness, and user experience are used. While these phrases make some sense at a surface level, once you start to examine them more closely, they become increasingly unsatisfactory as design requirements. For example, there is no universally accepted definition of what constitutes a user-friendly or intuitive design.

Attempts have been made to provide more scientific rigor to the issue of designing a product or system for use by people. Card, Moran, and Newell (1983) sought to deconstruct user interactions with computers into a series of micro-interactions. These micro-interactions can be analyzed and quantified.

Once quantified, they can then be optimized to ensure the most efficient interaction between the user and the computer. Card, Moran, and Newell describe a number of models of interaction derived from cognitive theory. These range from the simple model human processor, which describes interactions as the linear sum of perception, cognition, and motor cycles, to the more complex GOMS (Goals, Operators, Methods and Selector rules) model that attempts to classify higher-level generic types of interaction.

Such models are attractive in that they provide hard numbers with which to work and thus optimize a design. However, where they generally are weak is in capturing the essence of human unpredictability. For example, people do not always behave the same way when they are fatigued as when they are alert. Nor do they always choose the optimal path to solving a problem. Sometimes they are unsure what to do. While Card, Moran, and Newell's approach provides for optimizing the ideal path to a solution, it does not provide guidance on how to help users reach the stage where they are proficient enough with the technology to make the correct choices all of the time.

The field of usability engineering, as described so lucidly by Nielsen in his famous 1993 book (Nielsen 1993), takes a somewhat different approach to designing for people. Rather than trying to construct models of interaction from the micro-interactions up to a bigger picture, Nielsen first focuses on gathering interaction data from users and then trying to deconstruct it into types of behavior. He is particularly interested in the areas where the users either make mistakes or show signs of confusion. Again, there is little by way of guidance for how to design better, more user-friendly solutions as the focus is instead on finding problems.

Authors such as Norman (1988), Cooper (1999), and Shneiderman (Shneiderman and Plaisant 2010) have attempted to provide a narrative to how designers should think when approaching designing a product or system for use by

people. The typical approach is to use exemplar case studies to illustrate the principles being proposed.

A review of the literature shows many different approaches and no single, uniform, "best practice" approach to designing for people. The closest to a conformity of opinion in this area is that a user-centered design approach is the most reliable option for generating a usable, user-friendly, intuitive, etc., design.

What we can see from the above is that designing for people is a challenge, but one that designers face on a daily basis. In an attempt to minimize the variability in the design requirements and specifications, designers will typically try to reduce the users to a simple homogeneous representative. As Cooper warns (Cooper 1999), unless they are presented with a very specific view of who the users are, the designers will often substitute themselves as the target users. This is a seductive assumption to make. After all, designers are people and the users are people, so who is to say that the designers are not suitably representative of the target users?

Of course, taking a quick step back to view the problem and the problem with that assumption becomes clear. For example, not all users, and indeed very few users, will have the insights into the operation of the product or system that the designer will have. Consequently, the designers will be power users, whereas most target users will not. A single designer is also, by definition, homogeneous. By designing for himself or herself, the designer is making the assumption that all users will share his or her knowledge and also his or her physical attributes and functional capabilities. This raises the question of how valid is that assumption? Furthermore, if the assumption is invalid, how can the designers be assisted in designing products or services that better meet the needs of the wider population?

The solution is the principle of designing for inclusivity, i.e., designing so that as many people as possible can use the product or service being developed. The practice of doing this is known as inclusive design.

## Why Do We Need to Design More Inclusively?

The global population is not only increasing in size, it is also changing demographically. The population in many countries is no longer dominated by young adults, but is increasingly ageing. Indeed, countries such as Japan can be considered as having an aged population, with a median age of 41 years old (United Nations 2001). The current population ageing is unprecedented. Not only are people living longer, but the proportion of people under 15 is simultaneously decreasing. It is predicted that by 2050, approximately 21 % of the global population will be over 60 years old, representing an estimated two billion people. As a comparison, in 1950, the proportion was only 8 % of the population being over 60 years old. In more developed countries, the proportion of people aged over 60 will increase from just under 20 % in 2000 to 33 % of the population in 2050. Even the older population itself is ageing, with the fastest growing age group being the so-called oldest old, i.e., people over the age of 80. Finally, the potential support ratio, i.e., the number of younger adults available globally to support older adults is predicted to

fall from 12 younger adults per older adult in 1950 to only 4 by 2050 (United Nations 2001).

Since there are comparatively few working designers over the age of 60 and even fewer over the age of 80, the assumption that designers designing for themselves will lead to an acceptable design looks increasingly tenuous in view of the clearly established change in demographics being seen globally.

Such demographic changes are important and will have profound implications for designers, organizations, and service providers. In most developed countries, older adults typically have the highest disposable income – the so-called grey dollar or grey pound. They also tend to be more politically active and more likely to vote – "the grey vote" (Keates 2006). They also are more likely to want to live in their own home rather than moving in to live with their children or into sheltered accommodation. This desire to stay in their own home is partly driven by a strong sense of independence, but also a desire to ensure that they can leave a legacy for their children. For many such adults, their home represents a substantial inheritance that they wish to pass on to their offspring.

Furthermore, older adults typically display different attitudes to technology adoption than younger adults. For example, a report by the UK Office for National Statistics has shown that Internet usage declines with age. 99 % of all UK adults aged 16–24 years old had used the Internet, whereas only 34 % of all such adults over 75 years of age had done so (ONS 2013). This equates to 3.1 million people over the age of 75 who have never used the Internet in the UK.

The ageing of the global and national populations is associated with an increase in the prevalence of disability. The WHO defines disabilities as:

> . . . an umbrella term, covering impairments, activity limitations, and participation restrictions. An impairment is a problem in body function or structure; an activity limitation is a difficulty encountered by an individual in executing a task or action; while a participation restriction is a problem experienced by an individual in involvement in life situations.
>
> Disability is thus not just a health problem. It is a complex phenomenon, reflecting the interaction between features of a person's body and features of the society in which he or she lives. Overcoming the difficulties faced by people with disabilities requires interventions to remove environmental and social barriers. (WHO 2013)

For many people, a disabled person is typified by either a young man in a wheelchair or an older blind man walking with either a white cane or a guide dog. However, these are both anachronistic stereotypes that do not reflect the true variety or prevalence of functional impairments across the whole population. For designers, it is functional impairment – i.e., a limitation in someone's capabilities – that is important. Disability is a consequence of a person's functional impairments preventing them from interacting with a product or service successfully within a given context. If the product or service is designed to be sufficiently robust to support or accommodate a wide enough range of functional impairments or limitations, then no disability or handicap should be experienced by that person.

Disability, or more correctly functional impairments, is surprisingly prevalent across the population. For example, in 1996/1997, it was estimated that 17.8 % of

the population of Great Britain (i.e., England, Wales, and Scotland) had at least one functional impairment (Grundy et al 1999). These impairments could be further broken down into:

- 14.7 % having a motor impairment, such as locomotion, dexterity, reach and stretch, and strength
- 5.7 % having a cognitive impairment, such as difficulty with memory, recall, recognition, understanding, and communication
- 8.7 % having a sensory impairment, such as vision or hearing

It can be seen that 14.7 % + 5.7 % + 8.7 % >17.8 %. This inequality arises because approximately 8.8 % of the population has more than one class of functional impairment, e.g., a motor and a cognitive impairment. 2.5 % of the population has all three classes of impairment, i.e., motor, cognitive, and sensory.

This overall prevalence pattern is believed to be typical of the developed world. The US Census Bureau's 1999–2004 American Community Survey (ACSO 2007) asked respondents if they had any kind of disability, defined here as "a long-lasting sensory, physical, mental, or emotional condition." The data collected were as follows:

- 16.0 % reported any type of disability (cf. 17.8 % from the British survey).
- 4.7 % reported a sensory disability (cf. 8.7 % for a sensory impairment).
- 10.6 % reported a physical disability (cf. 14.7 % for a motor impairment).
- 5.2 % reported a mental disability (cf. 5.7 % for a cognitive impairment).
- 3.1 % reported a self-care disability (no direct comparison available).
- 4.9 % reported a go-outside-home disability (no direct comparison available).
- 5.6 % reported an employment disability (no direct comparison available).

Again, it can be seen that multiple impairments/disabilities are common. From the same survey:

- 6.7 % reported one type of disability (cf. 5.9 %).
- 7.6 % reported two or more types of disability (cf. 8.8 %).

Some of the differences in prevalence will arise from the different definitions used for each survey. However, the overall pattern is sufficiently similar for it to be assumed that approximately one in six adults in a developed country will have a functional impairment and approximately half of those people will have two or more classes of impairment types. Consequently, any mass market product or service can expect one in six of its target users to have at least one functional impairment.

This is a powerful argument for why designing for inclusivity is important. No company would like to potentially exclude 18 % of its target users. It is also worth noting that prevalence of disability typically increases with age (Christensen et al 2009).

Although older adults are increasingly healthy compared with their predecessors, the ageing process is still accompanied by an overall decrease in functional capabilities. Typically, several capabilities will degrade over time, and this leads to the widespread prevalence of multiple impairment classes. Someone who is older could easily have arthritis and cataracts and be a little hard of hearing, for example.

More recent evidence is also emerging that designing for inclusivity not only helps organizations gain access to market share by enabling potential customers who could not previously access a product or service to do so but also improves the quality of the interaction for other users. In industries where there can be high churn rate, such as for television service providers, there is evidence from Scandinavia that improving accessibility reduces the rate of customer churn substantially. Effectively, customers get used to a better level of service and experience from one particular provider and thus become more reluctant to swap to a rival (Looms 2011).

There is a societal impact to the prevalence of impairments as well. For example, a 2011 study by the UK Office of National Statistics has shown 45.6 % of adults of working age with a disability are in employment. This means that of the 7.1 million adults with a disability in the UK in 2011, only 3.2 million were in employment. Furthermore, those in employment were more likely to have a part-time job (33.8 %) compared with working people without a disability (24.6 %) (ONS 2011).

Given the ageing population and the decrease in the proportion of potential support ratio discussed earlier, it is essential that as many people are offered the opportunity of employment. It is to no one's benefit for over 50 % of any demographic sector to be excluded from employment.

Governments around the world have been increasingly taking note of such issues. Legislation has been passed in many countries outlawing discrimination in many situations including employment. Examples of such legislation include (Keates 2007):

- 1990 US Americans with Disabilities Act (ADA 1990)
- 1995 UK Disability Discrimination Act (DDA 1995)
- 1985 Canadian Human Rights Act (CHRA 1985)
- 1992 Australian Disability Discrimination Act (DDA 1992)

Perhaps the most original and innovative piece of legislation is Section 508 of the 1973 Rehabilitation Act (Sec508 1998). This legislation prohibits the US federal government and all of its agencies from purchasing, using, maintaining, or developing any electronic and information technology products that are not deemed fully accessible. Since the US government is currently the world's biggest purchaser of IT products, very few organizations in the IT arena can afford to choose to ignore such a huge market. Consequently, in the USA at least, the accessibility requirements specified in Section 508 have become de facto standards for the IT industry. Those requirements have been codified into checklists that IT organizations must comply with. Furthermore, other governments around the globe,

at the national and local levels, have seen the success of Section 508 and are actively investigating their own local variants.

Consequently, from all the reasons discussed in this section, there are clear social, ethical, business, and legal reasons why designing for inclusivity is important. It is becoming increasingly unacceptable to casually exclude up to 20 % of the population by failing to adopt suitably inclusive design practices.

Functional capabilities and impairments are not the only possible causes of exclusion from being able to use a product. Physical attributes can also be an issue. Most anthropometric data typically covers from the 5th percentile to the 95th percentile. Designing precisely for such a range of attributes runs the risk of potentially excluding 10 % of the population. An example of such exclusion would be the available viewing angles for a touch screen kiosk, say. Alternatively, users can also be excluded because of their prior experience or knowledge – or lack thereof. However, for our purposes here, we will focus on capabilities and impairments. The basic methods discussed, though, are extendable to a wider range of potential user attributes such as these.

So why are not all products and services designed to be inclusive? Some of the reason is because designers are often not aware of the need to design more inclusively. Put simply, very few design briefs spell out the intended target users to include a range of user capabilities. Where designers are aware of the need to design inclusively, they are often not provided with either the correct tools or resources.

So how can designers be assisted to design more inclusively? There is no one-size-fits-all solution. In the remainder of this chapter, we will examine how designers can be assisted in creating more inclusive designs. First, though, it is necessary to understand the goal that is to be accomplished. Let us begin by considering what inclusive design is.

## Defining Inclusive Design

Inclusive design has been defined by the UK Department of Trade and Industry as a design goal whereby

> . . .designers ensure that their products and services address the needs of the widest possible audience. (DTI Foresight 2000)

Meanwhile, the Royal Society for the Encouragement of Arts, Manufactures and Commerce (RSA), defines inclusive design as being

> . . .about ensuring that environments, products, services and interfaces work for people of all ages and abilities. (Keates 2007)

This definition broadens the scope of inclusive design from being about products and services to including the design of environments. As such, this definition is closer to the US concept of Universal Design (Follette Story 2001), which originated from the need to create accessible buildings.

The UK Design Council has a further definition:

> Inclusive design is not a new genre of design, nor a separate specialism, but an approach to design in general and an element of business strategy that seeks to ensure that mainstream products, services and environments are accessible to the largest number of people. (Keates 2007)

The definition that is arguably closest to our purposes here, though, is from the British Standards Institution, which defines inclusive design as:

> The design of mainstream products and/or services that are accessible to, and usable by, as many people as reasonably possible ... without the need for special adaptation or specialised design. (BSI 2005)

This definition positions inclusive design as being about trying to make products and services as inclusive as possible without specialist adaptations. Those specialist adaptations would typically be products that would be classified as "assistive technology," in other words enabling products for people whose physical attributes or functional capabilities do not permit them to otherwise access and use the product or service.

The need for assistive technology typically arises where a product or service has been designed without due regard for the full range of possible users. This disregard leads to potential users being excluded from the use of the product or service. Design exclusion is defined by the British Standards Institution as the

> inability to use a product, service or facility, most commonly because the needs of people who experience motor, sensory and cognitive impairments have not been taken into account during the design process. (BSI 2005)

Based on this notion of design exclusion, one possible approach to designing for inclusivity is that of countering design exclusion, as proposed by Keates and Clarkson (2003), where products are analyzed systematically to identify the demands that they place upon the users and whether those demands can be reduced to make the product more inclusive.

In the remainder of this chapter, we will thus examine approaches to the twin goals of designing more inclusively and countering design exclusion.

## Principal Approaches to Designing for Inclusivity

As society in general has become more sensitive to the needs and rights of all people to live productive, self-determined, and independent lives, organizations, governments, and rights groups have all recognized that practices need to change to benefit the widest possible range of people. Governments have typically acted on this through legislation, as discussed earlier. However, organizations are then left with the challenge of actually developing the products and services to meet this goal.

Historically, designing for inclusivity has been taken to be synonymous with designing for the disabled. This is very much an outdated view, but one that still persists. The typical approach to designing for disability was that it was usually

driven by an individual designer's passion or personal circumstance. Usually, it would be motivated by either someone close to the designer who had been born with a severe medical condition, such as cerebral palsy, or else someone who had acquired a disability through a trauma, such as a vehicular accident. The resulting designs were often considered part of a "garden shed" heritage, because that was where many of them were stereotypically assumed to have been designed and built. They were typified by a comparatively unsophisticated appearance, sometimes looking quite Heath Robinson. However, that does not detract from the fact that many of them served very useful purposes and could be quite literally life changing for their users.

Such products were usually expensive to produce and typically met the needs of only a comparatively small number of potential users. A more inclusive approach to inclusive design was required.

Most approaches to inclusive design can be considered to fall under one of two principal categories: *top-down* or *bottom-up*. In a 1993 paper, Maria Benktzon proposed a model of the general population as a pyramid, as shown in Fig. 1 (Benktzon 1993). The broad base represented the general population, with no significant functional impairments. However, the top of the pyramid represented those with very severe impairments and capability limitations.

Taking this user pyramid, a top-down approach to inclusive design is one that focuses on meeting the needs of the users at the top of the pyramid, i.e., those who are most severely impaired or limited. The assertion is made that if those users, i.e., those with the greatest constraints, can use the product or service, then so can the rest of the populations.

In contrast, the bottom-up approaches can be typified as taking an existing (mainstream) product or service, designed for use by the general population without necessarily any consideration for the widest possible range of users and then subsequently trying to make the product more inclusive. This increase in inclusivity is accomplished


**Fig. 1** The user pyramid (After Benktzon 1993)

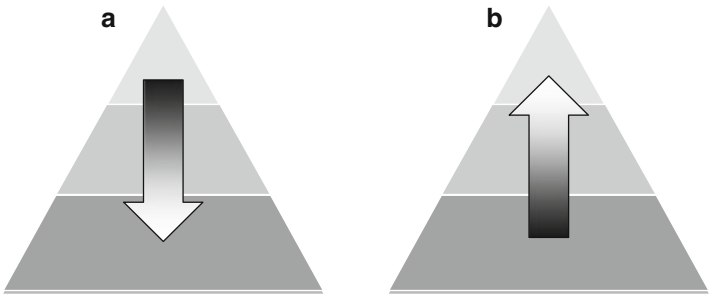



**Fig. 2** (**a**) The *top-down* approach. (**b**) The *bottom-up* approach

by trying to identify where users may encounter difficulties with using the product or service and then redesigning the problematic feature to make it easier to use.

Figure 2 shows the top-down and bottom-up approaches.

The fundamental concept of the top-down approach is straightforward to grasp. It can be distilled down to finding users with very severe functional impairments or capability limitations and design a product or service that they can use and then that is all. Everyone else should, in theory, be able to use the product. For example, if the product is designed for use by someone with severely limited vision, then everyone else with less limited vision should either be able to see everything of importance on the product or else be able to use their other senses (such as touch) to supplement the visual information. Typically, top-down design is accomplished through participatory design, i.e., recruiting users with severe impairments or limitations into the design team. They are often used for dual purposes, both to inspire the design and to test and evaluate it as the design evolves and develops.

The problem, though, is that the top-down concept begins to look a little strained the closer that you examine it. For instance, many users with severe functional impairments require solutions that would hamper other users. Think, for example, about the design of an ATM. A user in a wheelchair would want the keypad and screen to be at a lower height so that they can be reached and seen respectively from a seated position. However, someone with a visual impairment would want them both closer to their eye level so that they can be seen more easily. Similarly, designing a product or service with Braille on it only benefits those who can read Braille, which is a tiny proportion of the population – increasingly even blind people do not use Braille, preferring electronic communication systems, such as text to speech (TTS).

It is also impossible to recruit or study users with every possible functional impairment and so the products developed from a top-down approach necessarily end up being tailored to those users who are involved in the design process.

It is often considered that the products developed under a top-down approach are also typically more expensive. They can be thought of as custom-built solution to very specific issues and are commonly regarded as niche products. This is not always the case, though. A famous exception to this is the British Telecom (BT) big button telephone (Keates and Clarkson 2003), which was designed expressly for

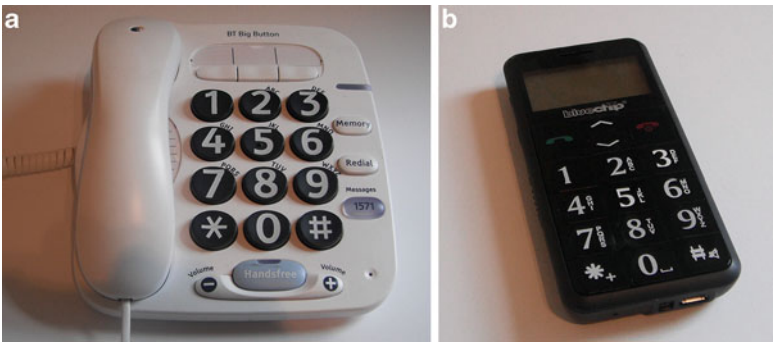



**Fig. 3** (**a**) The big button telephone. (**b**) A big button mobile telephone

older adults. As can be seen in Fig. 3a, the telephone has large buttons with high-contrast numerals on them. The buttons had very positive actions, with pronounced kinesthetic feedback when a button was depressed. The big button telephone has been on sale in various forms since the late 1980s and has been BT's most profitable phone since the days of rotary telephones. Similarly, the success of the OXO Good Grips brand has been based almost entirely on a combination of good quality kitchen utensils and tools married to an easy-to-grip handle design that was originally developed for users with difficulties gripping objects (Keates and Clarkson 2003).

The big button telephones design was made possible because desktop telephones do not have particularly tight size restrictions placed upon their design. Mobile phones, however, must be designed to be small and light. Consequently, the large big button telephone buttons simply cannot fit onto a mobile phones form factor. Compromises, such as shown in Fig. 3b, have been attempted, but at the cost of reduced screen size.

The bottom-up approach to inclusive design is generally the preferred approach for many organizations. There are many examples in everyday life where it is clear that a design has been changed slightly to make it more inclusive. For instance, the addition of Braille on new signs in a building or ATMs at different heights for wheelchair users on a wall outside a bank are examples where a design or product has been changed very slight to remove an accessibility issue. BS7000-6, and indeed most of the definitions of inclusive design, appears to tacitly support this notion, by making reference to the "mainstream" products and services being made more inclusive.

Bottom-up approaches can be implemented in a straightforward fashion for existing products. Simply recruit some users for a user observation study and watch them attempt to use the product. Note any difficulties that they have and then redesign the product to remove or lessen those difficulties. However, as for the top-down approach, things are not always so straightforward.

For example, which users should be recruited? What impairments or capability limitations should they have? How severe should those impairments or

limitations be? These are all nontrivial questions. Arguably, though, the most important question is how inclusive do we need to make the product?

These are all very important questions and transcend the fundamental top-down or bottom-up dichotomy. Let us look at these questions in detail:

- **What user impairments or capability limitations?** This question is comparatively straightforward to answer. The most systematic approach to answering it is to break down the interaction with the product into its most basic component steps and then list the perceptual (sensory), cognitive, and motor skills required for each step. This process will yield a list of capabilities required to interact with the product or service – and that will form the basis of what impairments and capability limitations to be considered.
- **How severe should those user impairments or limitations be?** If you are adopting a top-down approach, then users with the most severe impairments who would be expected to be able to use the final design should be recruited. For a bottom-up approach, it is arguably more helpful to recruit users who are at or are just beyond able to use the product, the so-called edge cases. The reality, though, is that in many cases the answer actually adopted is, "Whoever we can find," although this is clearly far from ideal.
- **How inclusive do we need to make the product?** Where a design is being driven by fear of legislation, there is a very real question of how inclusive the product needs to be to avoid ending up on the wrong end of a court action. Most of the legislation is framed in terms of "reasonable" accommodation, but what is "reasonable" is difficult to quantify unambiguously. Very few court cases brought under the Americans with Disabilities Act or the UK Disability Discrimination Act are resolved in court with a clear and unambiguous statement of when a design is sufficiently accessible or is otherwise illegally inaccessible. The majority of such disputes are often resolved via out-of-court settlements and the subjects of nondisclosure and confidentiality agreements. Implementation of Section 508 has led to the development of checklists for suppliers to check their products and services against, but such checklists are far from ubiquitous in all product and service domains. Organizations are typically left guessing as to what is legal and what is illegal. The safest option is to establish current best practice and to adhere as closely to that as possible, such as the Web Accessibility Initiative guidelines for developing accessible web pages (http://www.w3.org/WAI/).

The question of how inclusive a product needs to be has a further consideration, which is whether inclusive is simply about accessibility or about equivalency of experience. A simplistic view of inclusivity is that any design that provides access to functionality for an intended user is an inclusive product. However, user groups, such as major US charities are taking the view that this is not satisfactory. Their position is that a genuinely inclusive product must offer an equivalent experience for someone who has a functional impairment. For example, if it takes a person with unimpaired vision 10 min to complete a task, such as purchasing a ticket via an online web shop, then it should take someone who is blind, say, the same amount of

time to complete the same task. This argument builds on the US Americans with Disabilities Act being enacted as civil rights legislation, thus focusing on no discrimination in experience as a key goal.

There are hybrid approaches to inclusive design that are based around collected design best practice, accepted rules of thumb, or other design approaches from complementary fields such as ergonomics and human factors (e.g., Erlandson 2008). As these hybrid approaches are usually collections of recommendations, the individual recommendations may variously have come from the top-down or bottom-up approaches. As such, they may help designers provide some coverage of both schools of thought. However, care must be taken to ensure that the coverage provided is sufficiently comprehensive. Where approaches from other disciplines are suggested, these may need to be adapted to the wider set of users typically being considered for inclusive design.

A final consideration for organizations looking to produce inclusively designed products is how these should be marketed. BS 7000-6 argues that ideally integrated ranges of products that are suitable for all potential users should be created (BSI 2005). However, this may not always be possible. Consequently, a range strategy is required to choose between the following options:

(a) A complete integrated range without the need for adaptive accessories
(b) New models to be added to a range and adaptive accessories developed for existing products
(c) A complementary range, coordinated visually and technically to some degree with the existing range
(d) A separate range with no connection to mainstream offers (e.g., manufactured by a different company, marketed under a different brand, and distributed through a different supply chain)

Having looked at the different principal approaches to inclusive design, it is time to look at the pragmatic issues of implementing it.

## Implementing Inclusive Design

Let us begin by considering a generic approach to inclusive design. As discussed earlier, the most common approach to implementing inclusive design is to adopt a participatory design approach, recruiting users into the design team and having them serve as the inspiration for the designers. This is the basis taken for the Design Business Association (DBA) Inclusive Design Challenges, organized in conjunction with the Helen Hamlyn Centre (Cassim 2004). DBA member consultancies from all design disciplines are set a design challenge to create a mainstream product or service that can be enjoyed equally by users of all abilities. The teams work with the Helen Hamlyn Centre, disabled users, and other experts to ensure that all aspects of inclusivity are considered throughout the challenge. Prizes are awarded at the end of each challenge.

While participatory design approaches do, in general, lead to more inclusive designs when appropriate users have been selected, it can be difficult to find and recruit the users. Additionally, they are only one type of user-centered design practices. ISO 9241 Part 210 defines user-centered design as a subset of human-centered design, where the latter is defined as an

> . . .approach to systems design and development that aims to make interactive systems more usable by focusing on the use of the system and applying human factors/ergonomics and usability knowledge and techniques. (ISO 2010)

The ISO standard describes six key principles that will ensure a design is user centered:

- The design is based upon an explicit understanding of users, tasks, and environments.
- Users are involved throughout design and development.
- The design is driven and refined by user-centered evaluation.
- The process is iterative.
- The design addresses the whole user experience.
- The design team includes multidisciplinary skills and perspectives.

These are the same principles that need to be adhered to for inclusive design. They key feature of inclusive design, that separates is slightly from most user-centered design in practice, is the recalibration of who the users are. In effect, inclusive design is user-centered, or human-centered, design where the users are defined to explicitly include users with functional impairments or capability limitations.

The principal methods most commonly associated with user-centered design are participatory design, cooperative design, and contextual design. Participatory design and cooperative design are largely very similar, with the principal difference being that participatory design is more widely used in Europe and cooperative design in the USA. However, they both focus on the role of users as equal members within the design team. Contextual design arose from ethnographic methods, building up a better understanding of who the users are and the wider context of use of the product to be designed (Beyer and Holtzblatt 1998).

Where a wholly new design for a product or service is required, any of these user-centered approaches are valid and, if implemented correctly with an appropriate choice of users, should yield an accessible and inclusive design. However, these methods are expensive and time-consuming to adopt and so organizations are typically looking for cheaper and faster options for achieving similar outcomes. Such options can include:

- **Empathy**. Empathizing with the users is a very cheap method for implementing user-centered design. Put simply, the designers try to picture themselves as the users and cognitively walk through the process of using a

product or service from that user's perspective. Indeed, contextual design could be considered a highly developed variant of this. The effectiveness of this approach can be enhanced through videos, photographs, multimedia stories, etc. of the users. There are numerous websites available that offer invaluable information to support this approach, such as the inclusive design toolkit (http://www.inclusivedesigntoolkit.com/). The principal disadvantage to this approach is how it is calibrated. In other words, how do the designers ensure that what they think is the case, really is the case? That calibration needs to happen via supplementary or complementary methods, such as user evaluation/observation sessions.

- **User evaluation or user observation sessions.** Rather than recruiting users for the entirety of the design process, organizations may opt to recruit them for intensive sessions where the users interact with the product or service. The observations are then used in the redesigning of features that were problematic for the users. However, further iterations of user sessions are then required to ensure that the redesigned product is genuinely an improvement. One common approach is a frequent cycle of user evaluations at defined gateways in the design and development process. However, there is also a balance to be reached regarding the frequency of such sessions. Too frequent and the process becomes expensive and time-consuming. Too infrequent and the design being evaluated may have acquired more inaccessible features as the design will have progressed substantially in the intervening time period. The user sessions are required to keep the design "on track."
- **Simulation aids**. The physical aspects of particular impairments or capability limitations can be simulated through the use of aids, such as thick gloves (loss of feeling), ear defenders (hearing impairment), and a blindfold (blindness). Such aids can help identify many basic accessibility issues quite quickly. Attempts have been made to model user performance and behavior characteristics (e.g., Keates et al 2002), and these feed into simulation aid design (e.g., Biswas and Robinson 2008). However, it must be noted that it is possible to read too much into simulation, and both designers and researchers have been known to regard simulation as reproducing the entire experience of what it is to have a particular impairment instead of recognizing that only one aspect of the impairment is actually being simulated.
- **Outsourcing**. An increasingly common solution is to simply outsource the design to external experts in inclusive design. The level of outsourcing can vary from simple **expert assessment**, where an inclusive design expert performs an evaluation of the product or service, to a comprehensive design service, which encompasses all aspects of the design process. There is an increasing number of specialist agencies and consultancies offering such services to organizations, and this may be a particularly attractive solution for smaller organizations that may not have the necessary skills in-house and where acquiring them may be considered prohibitively expensive. The disadvantages to this approach include never developing the skills in-house and relying solely on the quality of the external agency or consultancy used.

- **Best practice/design guidance**. Arguably one of the most cost-effective solutions for generating at least a basically accessible design is to find design best practice recommendations and follow there. Examples include design for users with cognitive impairments (Keates et al 2007) and also design for managing inclusive design (BSI 2005). Such guidance needs to be followed with a degree of caution, though. It is necessary to ensure that the contexts of the guidance and the design process are sufficiently similar. For example, recommendations on absolute button size for a smartphone may not be the best choice for a touch screen kiosk. However, where the contexts of the guidance and the product being design are sufficiently similar, such guidance can be extremely effective in aiding the design and development of more inclusive products and systems.

It is obviously up to the company commissioning the design to consider the complexity of the design and the in-house skills and knowledge available when deciding which of the above approaches is most suitable to their needs.

The outcome of the above approaches will typically be a list of issues or problems to be addressed with the design, as would usually be expected following a user evaluation session, for example. Such lists can themselves be problematic for designers. It is often hard to prioritize which issues are the most important to fix first and, occasionally, which ones may actually harm the overall usability and accessibility of the product. This is difficult enough for designers where the users are homogeneous. In the case of inclusive design, though, they are often very heterogeneous. Consequently, organizations and designers need assistance to help prioritize the most important issues.

One very useful source of information is quantified data about the extent of exclusion caused by each item on the list of issues to be fixed. Effectively, each source of exclusion can be ranked by how many people are excluded by that particular problem, e.g., how many people cannot see that label and how many people cannot hear that beeper. Exclusion calculators, based on the data collected for the UK 1996/1997 Disability Follow-Up Survey are available at:

- http://www.inclusivedesigntoolkit.com/
- http://www.eng.cam.ac.uk/inclusivedesign/

These calculators allow designers to specify the capability demands placed on the users by each stage of the interaction with the product. The calculators then display how many adults within the population of Great Britain cannot meet those demands and thus would be excluded from using the product. Worked examples of how to use the calculators have been provided by Keates and Clarkson (2002, 2003).

Once difficulties with using the product or service are found and prioritized, it is recommended that designers and organizations consider the following options (Keates 2007):

- **Can this feature be removed?** If a feature is not essential to the operation of a product, then a discussion should be had over whether the feature should be retained if it is shown to be causing users difficulties.
- **Can this feature be changed to make it more accessible?** If a button is too small to be seen or is difficult to press, then the designers should consider making it larger.
- **Can a complementary method of offering the functionality be added?** If a user cannot hear the beeper that indicates the end of a washing machine cycle, then the designers should consider adding a countdown timer or a flashing signal light to complement the beeper.
- **Can the functionality be offered in an alternative way?** For example, if a button is causing difficulty, could a slider be used instead?
- **Can an auxiliary aid (or assistive technology) be offered to supplement to feature?** Televisions used to be controlled only by buttons on the television set itself. Nowadays, almost all televisions are controlled by remote controls, but still retain the on-set buttons. In effect, the control is now handled by a separate device. Many special purpose adaptations exist in the assistive technology domain, but their success depends on how well they couple with the product to be used. Where the product or service manufacturer makes both the main product and the assistive technology, this should not be a major issue. However, where the assistive technology is a third-party product, then care needs to be taken to ensure full compatibility between the technologies being used.

Building on all of the above, Keates and Clarkson (2003) proposed a seven-level model for inclusive design especially for use with interactive products, which was later adopted in BS7000-9 (BSI 2005). The model is as follows:

- **Level 1: Identifying user wants and aspirations.** Defining and then verifying the complete problem, including social acceptability requirements (i.e., would a user want to use this)
- **Level 2: Determining user needs.** Specifying the functionality to be provided and then verifying the functional specification, i.e., the functional acceptability
- **Level 3: Facilitating user perceptions.** Introducing appropriate output/feedback mechanisms and then verifying that users can perceive (i.e., see, hear, etc.) the output from the product
- **Level 4: Ensuring users understand how to use the product.** Structuring interactions that match user expectations of how the product should behave and then verifying that users understand the product's behavior and current state
- **Level 5: Ensuring users can interact physically with the product.** Developing quality of control and user input and then verifying that the users can control the product without undue physical discomfort
- **Level 6: Verifying that the product does what is intended.** Evaluating the total product's functionality, usability, and accessibility and then validating its practical acceptability

- **Level 7: Confirming that users are happy with the product.** Evaluating match with user requirements and then validating social acceptability

This is a flexible model that is intended to be applied iteratively throughout an overall design process. Its most important feature is the explicit focus on the user's perceptual, cognitive, and motor capabilities and how those map to the demands required to interact with the product or service being designed.

## The Future of Inclusive Design

As products and services become increasingly "smart" and flexible, the days of a single interface for a particular product or service are numbered. For example, a kettle that is connected to a smart home environment no longer needs to be switched on via a switch on the body of the kettle. The smart homes of the future will have computing power built in to almost all of the gadgets and appliances in the house. Such systems could be controlled through mobile devices, such as iPads, with applications for controlling particular gadgets installed within a universal framework. In such a scenario, the kettle could be switched on via a command sent from such an iPad through the home network to a receiver in the kettle.

Moving interfaces from the physical world to the virtual world means that many of the physical constraints, such as size and weight, that currently restrict design options may no longer apply. The switch for the kettle could be as big as the screen of the iPad without affecting the size and weight of the kettle itself. Furthermore, the interface itself could evolve and develop. Evolutionary user interfaces (EUIs) are being design in research labs where the interfaces become more complex and powerful as the user's skill and knowledge increases. So, a novice user is offered a simple interface with limited options, but a lot of help and guidance. As the user becomes more proficient, more options and features are introduced, and the guidance is simultaneously reduced until the user is fully competent in controlling all aspects of the device.

The basic concept of EUIs has been around for some time, especially in the computer games industry, where users are often trained in increasing levels of complexity of game control as they progress through multiple levels. Games have the advantage, though, of knowing the full user context at any point in time, whereas EUIs do not nor, consequently, what level of difficulty the user should be ready for.

An EUI needs a context-agnostic approach to establishing the level of a user's skill in order to select the most appropriate level of "difficulty." One approach to this is to develop a method of automated skill assessment where users complete a set of known tasks and their ability is computed from the classified results of those tasks. EUIs are most likely still some time away from becoming commonplace, but they are one possible option for more inclusive products in the future.

## References

ACSO (2007) American Community Survey 2004 subject definitions. American Community Survey Office, US Census Bureau

ADA (1990) Americans with Disabilities Act (US Public Law 101-336)

Benktzon M (1993) Designing for our future selves: the Swedish experience. Appl Ergon 24(1):19–27

Beyer H, Holtzblatt K (1998) Contextual design: defining customer-centered systems. Morgan Kaufmann, San Francisco

Biswas P, Robinson P (2008) Automatic evaluation of assistive interfaces. In: Proceedings of the 13th ACM international conference on intelligent user interfaces, pp 247–256

BSI (2005) BS 7000-6:2005 design management systems. Managing inclusive design. Guide. British Standards Institution, London

Card SK, Moran TP, Newell A (1983) The psychology of human-computer interaction. Lawrence Erlbaum, Hillsdale

Cassim J (2004) Cross-market product and service innovation – the DBA challenge example. In: Keates S et al (eds) Designing a more inclusive world. Springer, London, pp 11–19

CHRA (1985) Canadian Human Rights Act (R.S. 1985, c. H-6)

Christensen K, Doblhammer G, Rau R, Vaupel JW (2009) Ageing populations: the challenges ahead. Lancet 374(9696):1196–1208

Clarkson PJ, Keates S (2002) Quantifying design exclusion. In: Keates S, Clarkson PJ, Langdon PM, Robinson P (eds) Universal access and assistive technology. Springer, London, pp 23–32

Cooper A (1999) The inmates are running the asylum. SAMS Publishing, Indianapolis

DDA (1992) Disability Discrimination Act 1992 (Australia)

DDA (1995) Disability Discrimination Act 1995 (c. 50)

DTI (2000) A study on the difficulties disabled people have when using everyday consumer products. Government Consumer Safety Research, Department of Trade and Industry, London

Erlandson RF (2008) Universal and accessible design for products, services, and processes. CRC Press, Boca Raton

Follette Story M (2001) The principles of universal design. In: Preiser W, Ostroff E (eds) Universal design handbook. McGraw-Hill, New York

Grundy E, Ahlburg D, Ali M, Breeze E, Sloggett A (1999) Disability in Great Britain: results from the 1996/7 disability follow-up to the Family Resources Survey. Charlesworth Group, Huddersfield

ISO (2010) ISO 9241–210:2010 Ergonomics of human-system interaction – part 210: human-centred design for interactive systems. International Standards Organisation, Geneva

Keates S (2007) Designing for accessibility: a business guide to countering design exclusion. Lawrence Erlbaum, Mahwah

Keates S, Clarkson PJ (2003) Countering design exclusion: an introduction to inclusive design. Springer, London

Keates S, Langdon P, Clarkson PJ, Robinson P (2002) User models and user physical capability. User Modeling and User-Adapted Interaction (UMUAI). Wolters Kluwer 12(2–3):139–169

Keates S, Adams R, Bodine C, Czaja S, Gordon W, Gregor P, Hacker E, Hanson V, Kemp J, Laff M, Lewis C, Pieper M, Richards J, Rose D, Savidis A, Schultz G, Snayd P, Trewin S, Varker P (2007) Cognitive and learning difficulties and how they affect access to IT systems. Int J Univers Access Inf Soc, Springer 5(4):329–339

Looms PO (2011) Making television accessible. International Telecommunication Union, Geneva

Nielsen J (1993) Usability engineering. Morgan Kaufman, San Francisco

Norman D (1988) The design of everyday things. The MIT Press, London/Cambridge, MA

ONS (2011) People with disabilities in the labour market, 2011. Office for National Statistics, London

ONS (2013) Internet access quarterly update, Q1 2013. Office for National Statistics, London

Sec508 (1998) Section 508 of the Rehabilitation Act (29 U.S.C. '794 d), as amended by the Workforce Investment Act of 1998 (US Public Law 105–220), 7 Aug 1998
Shneiderman B, Plaisant C (2010) Designing the user interface: strategies for effective human-computer interaction, 5th edn. Addison-Wesley, Reading
United Nations (2001) World population ageing: 1950–2050. Population Division. Department of Economic and Social Affairs. United Nations, New York
WHO (2013) Disabilities. World Health Organization. Available at http://www.who.int/topics/disabilities/en/. Accessed 10 July 2013

# Design for the Value of Presence

Caroline Nevejan and Frances Brazier

## Contents



**Abstract**

This chapter elaborates on design for the value of presence. As digital technologies have made it possible for us to connect to each other at a speed and scale that is unprecedented, presence is acquiring many new stances. The distinctions between being there (in virtual worlds), being here (making the being there available here), and the merging realities of these two are essential to the notion of presence. Understanding the essence of presence is the focus of current

C. Nevejan (✉) • F. Brazier
Delft University of Technology, Delft, The Netherlands
e-mail: nevejan@xs4all.nl; f.m.brazier@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_16

presence research to which many disciplines contribute, including computer science, artificial intelligence, artistic research, social science, and neurobiology.

The definition of presence used in this chapter is "steering towards well-being and survival," and this definition introduces a neurobiological perspective on presence fundamental to the approach on which this chapter focuses. This perspective recognizes the choices and trade-offs involved in presence design. Presence design is a meta-design, which creates the context for human experience to emerge. Presence as a value for design can be a design requirement, a factor of analysis, and a key value in a process of Design for Values.

This chapter discusses a number of analytical and design frameworks for constructing and deconstructing presence design. Acknowledging that presence is a fuzzy concept and that a variety of open issues can be identified, presence as a value for design is fundamental for human beings to accept responsibility in complex environments. Further research will need to address how we, as human beings, change and how our sense of presence changes, as a result of living in a network society with ubiquitous technology and all pervasive media being part of our day-to-day lives.

**Keywords**

Presence • Value • Design • Trust • Experience • Networks

## Introduction

Presence is a word that appears in many social, political, religious, and economic contexts and refers to an array of meanings. In the era of ubiquitous media, networks, and many complex infrastructures on which society depends, presence is no longer solely coupled to physical reality. Presence has acquired new virtual stances, with completely new dynamics. We, as human beings, connect to each other in many different ways. We meet virtually and participate in many different types of networks in merging on- and offline realities. We also participate in new types of communities such as energy communities in which participants organize their own exchange of energy. Energy communities rely on communication and visualization technology, but also on technology needed to provide data, for example, on usage, pricing, availability, accounting, and expected market developments mandating distributed data aggregation and service level agreements between participants.

To take responsibility we, as participants in such communities, need to have some form of presence for each other, both in on- and offline context as well as in information and communication trajectories. The design of presence is a prerequisite to participation: understanding the value of presence is a prerequisite to the design of large distributed complex participatory systems.

Human kind has been mediating presence since the beginning of times: leaving traces, making maps and drawings, telling stories, and performing rituals, music, and play. These are all ways with which we communicate presence from one time

or place to another, from one human being to another. Technology has made it possible for us to mediate our presence in new ways facilitating communication, interaction, and transactions over distance, often simultaneously. With the introduction of every new medium, new ways of establishing connection, and being able to say "hello," for example, is the first achievement and source of surprise and curiosity. Soon after, when many people start to use a new medium, this is integrated in day-to-day practices of millions of people and new habits, customs, and understanding emerge (Wyatt 2004). While new technologies produce increasingly better ways to produce mediated presence, natural presence is still distinct from mediated presence.

We, as human beings, are creative and find unexpected ways to survive and serve our own well-being. New technologies, new systems, are emerging continuously, connecting people across the world, creating connections between family friends and total strangers. These connections can be beneficial or detrimental for those involved. Facebook, for example, is designed to anticipate specific types of behavior with participatory scripts to build on this human potential of connecting with others. The outcomes of human behavior, however, cannot be predicted, and unintended side effects happen. The real-time connection between dozens, hundreds, and thousands of people Facebook provides has shown to be powerful for gathering people both for the good and for the bad. Social networks were instrumental to the rising of the Arab Spring between 2010 and 2012, to the hooligan gathering in London in 2011, and to Project X in Haren in the Netherlands in 2012.[1] In all of these events, the behavior of many individuals together creates a different situation and experience than any individual alone could have anticipated. In social networks individual behavior is contextualized and inspired, and this leads to new formation of (historical) experience, which is focus of further research in a variety of domains (Castells 2012).

This chapter elaborates on the notion that presence is essentially the strive for well-being and survival. Designing for the value of presence is not designing for a specific behavior. It is designing for experience, as argued in this chapter. Presence as a value for complex systems design has great societal relevance. Research into this value is timely.

## Explication of the Value of Presence

In today's ever changing network society, the amount of multimedia information we can access within seconds is unprecedented: we are, in fact, experiencing a tsunami of information at a speed that society has not experienced in the past. Our experience of time, place, and authenticity is changing (Benjamin 1936; McLuhan 1964; Baudrillard 1983; Postman 1985; Virilio 1989; Lovink 2012).

[1]Project X started off with a birthday invitation via Facebook and resulted in riots in which thousands of young people participated.

Some argue there are possibilities as never before; others claim that in the tsunami of copies at grand scale and speed, the concept of "meaning" implodes. In these times of fast transformation into the network society, place and time are still distinct factors in human lives and the social structures that are built. It is often, however, unclear how the "space of places" in the physical world relates to the "space of flows" in the many networks in which we participate (Giddens 1984; Castells 1996). In the collective experience of the emerging society, a new culture and a "next nature" is emerging in which we redefine, design, and establish how we want to live our lives (Mensvoort and Grievink 2012; Lunenfeld 2003). In the flow of images, text, and audiovisual communication, a new sense of authenticity is emerging creating media auras as a result (van der Meulen 2011). Key to this new culture and next nature is how we perform presence and participate in the complex networks that constitute our day-to-day reality (Brazier and Nevejan 2014).

The many online experiences and representations of selves mandate a new perspective on design of social, technical, and ecological networks and infrastructures, including consideration of related values such as privacy, integrity, and trust. The ethical dimension of presence design, including augmented reality design, is acknowledged as a value for the design of larger social technical and ecological infrastructures in a variety of public debates around privacy, integrity, and trust (Hamelink 2000).

As mentioned above, different notions of presence function in a variety of social, political, religious, spiritual, and ideological contexts. The focus of this chapter is on our natural presence qualified by breathing and a heart that ticks. It grounds presence in our physical nature.

## Existing Conceptualizations of Presence

Even though it was not labeled as such in a wide variety of scientific domains, presence research has been conducted over the last few centuries: in Philosophy, in Architecture, in Psychology, in scientific technology development, and in Communications and Media Studies. The distinction between being present in the here and now and being present elsewhere, by voice or by imagination (e.g., when reading a book), has been a topic of scientific interest for many years. The current large-scale spread of digital and distributed technologies has positioned the design of presence center stage.[2] With the ever developing technology, spreading Internet, evolving game culture, augmented reality, wearables, smart textiles, avatars, and more, new presence designs and configurations are continually influencing the possible

[2]With the rise of the network society, since the 1990s, notions of presence, tele-presence, mediated presence, and network participation were explored in many conferences like SIGGRAPH, CHI, Doors of Perception, ISEA, and Presence Conferences of the ISPR. The International Society of Presence research (ISPR) was founded in 2002 as a platform for international exchange.

stances of presence. The five key notions that have and still guide presence design during the last two decades are (1) being there, (2) being here, (3) merging realities, (4) presence as the strive towards well-being and survival, and (5) copresence, social presence, and witnessed presence. These notions and their historical context are discussed below in more detail.

### Being There

To create digital technologies for mediating presence, psychologists and computer scientists have been exploring mediation by the senses and the brain in relation to mediation by technology, in "(tele-)presence." Hundreds of experiments have been carried out to create and analyze the sense of presence in virtual environments. Different soft- and hardware applications have been created and studied to better understand how virtual experiences become real experiences for people involved. The target is to create the sense of "being there" (Lombard and Jones 2007). A typical experiment concerns the breakout of a fire in a virtual environment such as Starlab in Barcelona orchestrated. When people start to run away from a virtual fire, the sense of presence is high: these people are convincingly engaged in a situation of "being there" (Spanlang et al. 2007). As technology improves, VR is becoming a consumer product entering our homes and lives (Slater 2014).

Most studies on facilitating the sense of presence in virtual worlds explore our capacity of perception, attribution, imagination, and cognitive capacities when triggered or seduced by specific configurations of technology. Reliability, validity, sensitivity, robustness, non-intrusiveness, and convenience are criteria to which the literature refers (IJsselsteijn 2004; Hendrix and Barfield 1999). Both objective and subjective methodologies for measuring results have been developed (van Baren and IJsselsteijn 2004). Objective corroborative methodologies include psychophysiological measures, neural correlates, behavioral measures, and task performance measures. Subjective methodologies include (many) presence questionnaires, continuous assessment, qualitative measure, psychophysical measures, and subjective corroborative measures.

### Being Here

In the 1990s, the "being-here" perspective on presence design is initially overshadowed by the many commercial promises of technology to create time- and place-independent connections and communities. The possibilities of the new technologies are also, however, explored in less commercial settings aimed to contribute to local communities. Felsenstein's Community Memory project in San Francisco, the Domesday project in the UK, Geocities in the USA, and, for example, the Digital City of Amsterdam facilitate thousands of people to explore and co-design online experiences in the emerging digital culture at the time (Castells 2001). The quest in these initiatives was to create added value by using ICT technologies for local community involvement. The challenge was and is to make the "being there" of relevance to the "being here."

This is also the perspective taken by Gullstrom in which the influence of framing in architecture leads to the basis for new architectures for presence in which other

places through elaborate visual perspectives, with or without the use of technology, are made present as "being here" (Gullstrom 2010).[3]

In 2007, Nevejan claims there is a direct relation between design for presence and design for trust in the emerging network society in which on- and offline realities merge in which ultimately the "being here" is distinct (Nevejan 2007). Mediated presence contributes to language and concepts we as people share, but natural presence, the being here, is distinct because it holds the ethical dimension of an individual life. The physical steering towards well-being and survival is distinct for our individual lives and is distinct from how we touch each other's lives, as discussed below. In pain we respond different to our environment than when we are healthy and fit. When being in each other's physical presence, we can literally care for each other. When in conflict, physical presence allows for more expression in both aggression and compassion. Communication with others, who have other perceptions and convictions, has more bandwidth in natural presence than in mediated presence. This is a reason why project teams at the beginning and at the end of a project often come together in real life. Then they can ask "What is good to do?" and "Is it good what we do?"

## Merging Realities

In communication trajectories we incorporate on- and offline interaction into one experience over time. Buying an airplane ticket, checking in online, and boarding the plane physically offer an integral experience with a particular carrier. The easyJet experience, for example, is different from the Jet Airways experience. In personal relationships on- and offline moments create a specific communication trajectory that characterizes the experience of that particular relationship.

In 2005, Floridi proposes that local and remote spaces of observation and different levels of analysis define presence, given the complex dynamics between presence and absence (Floridi 2005).

Gamberini and Spagnoli extend the notion of tele-presence into a day-to-day experience of different simultaneous information and communication flows (Spagnoli and Gamberini 2004).

Since 2010, (tele-) presence in "traditional" virtual reality is studied in the context of cyber therapy. Focusing primarily on cognitive behavioral therapy, a deliberate

[3]Architect Gullstrom eloquently described 500 years of architecture history as a history of presence research in which elaborate processes of framing in different media format human presence and suggest other people, religious entities or other worlds, are being present here. Perspective and gaze, interaction, and attribution trigger the sense of presence. After analyzing buildings and paintings since the early 1600s, she describes in detail how since the 1970s in Palo Alto, in cybernetic circles, in the work of artists, and in many cultural events technology is used to create new architectures for presence in which other places are made present as "being here." As a result Gullstrom created an architectural "presence design toolbox" consisting of shared mediated gaze, spatial montage, framing and transparency, lateral and peripheral awareness, active spectatorship, and offscreen space.

bridge between the virtual and the real is created to synthesize in human experience events that are healing (Wiederhold 2006; Riva 2008).

In augmented reality the "being here" and the "being there" are presented in one interface. Virtual data are spatially overlaid on top of physical reality, providing the flexibility of virtual reality grounded in physical reality (Azuma 1997). Mediated reality refers to the ability to add to, subtract information from, or otherwise manipulate our perception of reality through the use of a wearable computer or handheld device (Mann and Barfield 2003). Current technology providing stereoscopic vision in shared augmented space, coupled to data repositories, merges these two realities. Recent results (Poelman et al. 2012) show the need to explicitly design mediated and witnessed presence for awareness and trust.

Some recent research on the affects of social networks can be described in terms of presence research into merging realities. For example, Danah Boyd studies how social networks affect teenagers' day-to-day life, actually revealing how what she calls "network publics" affect the performance of presence of these teens (Boyd 2014).

In the design of participatory systems, the concept of merging realities is embraced as a starting point of design. Focusing on the performance of presence in network contexts, in which on- and offline communication merge in our individual experience, new spaces for design unfold (Nevejan and Brazier 2010).

## The Strive for Survival and Well-being

Having identified that the sense of "being here" and the sense of "being there" are merging, the notion of presence needs to be (re-)considered. How can the essence of presence be formulated to include being there and being here in merging realities?

In 2004, inspired by the work of Antonio Damasio, Riva with Waterworth and Waterworth introduce a neurobiological perspective on presence (Riva et al. 2004) that does not depend on technology and allows for understanding presence in the context of merging realities. This neurobiological perspective on "presence" claims that the strive for well-being and survival, or what Spinoza referred to as "the conatus," is the essence of presence (Damasio 2004). Sensations, emotions, and feelings inform us of the direction in which well-being and survival can be found. We steer towards sensory sensations, emotions, and more complex feelings of solidarity, compassion, and love, and we steer away from pain, hate, and unpleasantness (Damasio 2000). We "perform" presence (Butler 1993). When touching a burning stove, we retreat immediately. When entering a place with a bad smell, we walk away. When meeting a big angry-looking man in a dark alley, we run. When an atmosphere suddenly turns into dispute and fights, we prefer to leave. And vice versa, when we see other people do good and nourish the sense of solidarity, we are inspired to do so as well.

Damasio also suggests that it is likely that the steering towards one's own survival and well-being includes the well-being and survival of others as well (Damasio 2004). Seeing pain of others hurts, aggressive behavior leads to unsafe situations and people will turn away. When transposing this suggestion to a network reality, new questions arise. Is it likely that when we think of mediated presence in which one does not have to confront physically the consequences of one's actions that an individual would

develop feelings of compassion or solidarity? How can consequences of our actions be felt in mediated presence? This is, for example, a major issue in training pilots using a flight simulator. Most of today's pilots have played with flight simulators in games in which the notion of "crashing" implies restarting the game.[4] Such considerations are related to the notion of presence as a value for design.

### Copresence, Social Presence, and Witnessed Presence

Individual performance of presence is affected and inspired by other people's presence. In 1963 Goffman introduced the notion of copresence, to refer to the situation in which we perceive others and in which we can sense that they perceive us (Goffman 1963). This research continues today. Researchers are still studying and measuring under what conditions copresence emerges in virtual environments and augmented reality applications and is accepted by people acting in these environments (Nowak and Biocca 2003).

In communications theory, social presence in social interaction using media and telecommunications refers to the differences in spheres of intimacy that a phone call or a face-to-face meeting, for example, generates (Short et al. 1976). Social presence is one of the pillars for educational design in blended learning contexts (Whiteside and Garret Dikkers 2012).

Copresence and social presence do not address the issue of the establishment of truth and trust, both fundamental to understanding what happens next in any social situation. Being and bearing witness to each other is historically the social structure in which truth and trust are negotiated. Nevejan argues that witnessed presence is fundamental for establishing trust both in the online and the offline world (Nevejan 2007; Nevejan and Gill 2012). An action that is witnessed becomes a deed. The witness can intervene in the course of events and can bear witness and testify which may change the understanding of the deed. Witnessing, as a way of having presence that includes the acceptance of responsibility for words and deeds, includes notions as addressability, response-ability, and clarity of subject positions (Oliver 2001). It also appears that to be witness includes to self-witness. The artistic research project Witnessing You concludes that "self-witnessing" is fundamental both to the process of being witness and the process of bearing witness (Nevejan 2012). The same conclusion is drawn in VR research into being there (Slater 2014).

## Main Issues of Controversy on the Notion of Presence as a Value for Design

A first issue of controversy is that presence is a fuzzy concept. Most measurements in the "being there" approach to presence design are concerned with effects of certain media configurations focusing on a reported sense of presence. Where does presence as a phenomenon start and where does it end? What is the opposite of

[4]Personal communication with military staff at Thales office in Delft in 2007.

presence? Not having presence may not be the same as being absent or having/performing absence. How can presence be defined to make distinctions possible between more or less, better or worse, real or false presence? The notion of witnessing sheds light on these issues, but does not make hard distinctions possible.

A second issue is the controversy of how presence is considered – as a result of human consciousness or as part of human consciousness. The notion of presence differs between the variety of social sciences and natural sciences, between deterministic and more holistic approaches. Also the role of emotions and the role of imaginations in processes of presence are approached differently. This regularly leads to misunderstandings.

A third issue concerns design trajectories of presence in complex systems. Inducing and deducing dynamics in virtual simulations and serious games require rigorous analytical skills and an associative/creative capacity at the same time. Results often only shed light on a specific dynamic given a set of predefined rules and variables. Nevertheless, these simulations and serious games inform real-life processes in which real people participate. The gap between simulations and serious games and real-life situations is considerable and has to be taken into account. In cyber therapy this gap is used to induce healing processes in individuals. When complex systems assist in matters of life and death, as in crisis management systems, or more mundane applications as in railway systems, unintended side effects can have dramatic effects. Virtual simulations and serious games are unable to anticipate how individuals will act and be witnessed in extreme situations in their strive for survival and well-being. Such unexpected side effects are matters of concern.

## What Does It Mean to Design for Presence?

Designing presence in complex systems in the context of the functionality and nonfunctional requirements on which a system is based should target specific functionality, such as to facilitate social interaction, to facilitate collaboration, to facilitate exchange, to facilitate a marketplace, and to facilitate distributed structures of governance. As the design of presence is not often explicitly addressed as an explicit requirement, it is often neglected. Developments in the outsourcing industry in India, for example, indicate that neglect for presence design is detrimental for the workers involved (Ilavarasan 2008; Upadhya 2008). Presence as a value for design, as a requirement, facilitates designs that make it possible for us to be able to have agency, accept responsibility, and be able to engage with others in meaningful interaction, making it possible for us to steer towards our own well-being and survival.

### Meta-design for Choices and Trade-Offs

Presence research is a science of trade-offs (IJsselsteijn 2004). We, as individuals, make these choices and trade-offs on the basis of what we know: we decide on how,

when, and where we perform our own presence in which situations. Collective experience with a medium affects how we, as a society, understand and respond to media realities. When film was just invented, and a train was approaching, the whole cinema audience would dive under the chairs. For many years email was ignored as a legitimate form of communication – it took up until a decade ago for the Courts of Law to accept email as proof. (Note that the concerns with respect to legitimacy of email are well founded.) The implication of understanding presence as a choice and trade-off, on both the individual and collective level, is that presence can be designed, and this opens up new fields for research and design.

There is a direct relation between design for presence and design for trust in the emerging network society in which on- and offline realities merge. Arguing that witnessed presence is fundamental for establishing trust, Nevejan (2007) introduces the YUTPA framework[5] in which four dimensions of time, place, action, and relation define potential trust in different presence configurations of these dimensions. Interdisciplinary research with artists, academics, and experts elaborated this framework and identified factors of significance in human experience in each dimension (Nevejan and Brazier 2012), providing a frame of reference for the analysis of choices and trade-offs in presence design.[6]

Designing for the making of choices and trade-offs, designing a context in which people can steer towards well-being and survival, needs to conceptualize presence design as meta-design (Fischer 2013). It is not designing for a specific behavior; it is designing for the choice of behavior or the creation of new behavior. Social networks, Internet platforms, and participatory systems aim to offer such meta-design upon which we can perform our own presence in our own way. Presence as value for design is mandatory in these systems of participation (Brazier and Nevejan 2014).

## Design for Experience

Design for presence needs to include the complex notion of design for experience. We make choices for our own behavior, for the performance of our presence, not only out of habit of previous behavior. Such choices are more complex and include outcomes of reflection on our previous action and outcomes, understanding of contexts, and imagination and anticipation of possibilities. Different levels of consciousness (proto, core, and extended) influence performance of presence (Damasio 2004).

In the English language, the word experience reflects different kinds of experience in one word only. In the German language, the word “erfahrung” is distinct

[5]YUTPA is acronym of “to be with You in Unity of Time, Place and Action”.

[6]This framework is fundamental to the analyses of human network interaction in the emerging participatory systems design paradigm that is studied and developed at Delft University of Technology (Brazier 2011).

from "erlebnis." A distinction is made between "erlebnis," referring to sensations and happenings, which are foundational to behavior, and "erfahrung" which refers to experience, as being the reflexive context in which we, as human beings, reflect upon our own actions and understand our own situation to inform new actions. Design for presence not only includes design for sensations and behavior ("erlebnis") as discussed above. Design for presence is distinct because it necessarily includes design for experience ("erfahrung") in which a larger context allows for individual reflection and choices. Performance of presence emerges from experience.

Experience design is a relatively young discipline in certain design schools in Europe, the USA, and India. Its theoretical foundation is diverse including media and cultural studies, marketing and business, philosophy, and interaction design.

Not often used today, but very clear in their intention, is the work of the Frankfurter Schule on experience design in the previous century (Habermas 1983; Negt and Kluge 1972). This group of German philosophers and social scientists posed the question of design for experience, as the ground for human's autonomous choice, in the early 1960s. Confronted with the fact that millions of people had followed Hitler in the 1930s and into WOII, they were determined to understand how individual people could keep their autonomy and independent perception in mass media and propaganda contexts. As result the Frankfurter Schule introduced a specific idea about experience design in which sensations and happenings need to be historically contextualized in both personal and collective ways to nurture reflection and inspire people to steer towards their own, and others, well-being and survival. Artists and artistic research play a role of significance in this approach. In the era of ubiquitous computing and all pervasive media, the thinking of the Frankfurter Schule is acquiring new attention.

## Artistic Research

Presence research uses many methodologies from the medical and natural sciences as well as methodologies from the social and design sciences. Artists, who have challenged the imagination of presence design with elaborate use of technology for several decades now, make specific contributions to presence research.

Every new technology is an inspiration for artists. They run with it, push its limits, and focus on exploring experiences that the new medium facilitates. For over 50 years now, technology artists have experimented with different presence designs. Using radio and television, video, audio, and digital media in many ways, artists have explored how human beings can perform presence in different media configurations. Marcel Duchamp, John Cage, Nam June Paik, Bill Viola, Char Davies, David Rokeby, Shu Lea Cheang, and Lisa Autogena, just to name a few, have altered the way in which people experience the merging realities around them.

Artists are experts in creating experiences for others offering perception and reflection in unanticipated ways and affect the aesthetic experience that is part of everyday life (Dewey 1934). Artistic research, including the making of work and methodologies for research, offers radical realism, non-conceptualism, and contingency (Schwab and Borgdorff 2014).[7] Distinct from art history, and distinct form art practice, artistic research aims to contribute to larger research questions (Biggs and Karlsson 2011; Borgdorff 2012; Zijlmans 2013). Presence design in the era of ubiquitous computing and pervasive media is definitely such a question.

## Different Analytical Frameworks for Constructing and Deconstructing Presence as Value for Design in Larger Social, Ecological, and Technological Structures

In a variety of disciplines, scholars are concerned with understanding requirements for designing structures in which we, as human beings, can steer towards our own well-being and survival for establishing sustainable social structures. None of these approaches are currently considered part of presence theory or design. However, when accepting that presence is essentially the strive for well-being and survival, these approaches contribute to presence design for larger social, ecological, and technological structures. Fundamental to all of these approaches is that we participate with our own strive for well-being and survival while participating in a larger process of collective evolution or change. Each of these approaches is concerned with design processes or analyses as meta-design for the value of presence.

### Business Studies: Presencing and the U-Turn

Senge[8] introduced the concept of "presencing" as a means to guide organizations to go through collective change (Senge et al. 2004). Presencing is defined as being aware of the here and now and imagining and anticipating what could happen next. In other words, "presencing" explores potential steering towards well-being and survival. With his colleagues Senge developed the concept of the U-Turn, in which an organization go through seven phases in which the "presencing" of its members is crucial in the dynamic to unfold change. The seven phases are expose, reorientate, letting go, emerge, crystalize, prototype, and institutionalize. Presencing is used as design requirement for organizational change.

[7]Currently the Society for Artistic Research hosts the Research Catalogue in which several journals on artistic research are published and debates are orchestrated.

[8]MIT Sloan School of Management, founder of the Society for Organizational Learning.

### Design Thinking: Collaborative Authoring of Outcomes

In complex design trajectories, in which business and political dynamics are at stake, the need to incorporate individuals strive for well-being and survival is acknowledged and has been studied in depth. To this end, Humphries and Jones[9] formulate the concept of "collaborative authoring of outcomes" (Humphries and Jones 2006). Through an iterative process of design in which different scenarios are explored by participants (stakeholders) in the to be designed new system/structure, the individual strive for well-being and survival drives the process of design. Because individuals participate from out their own strive for survival and well-being, and contribute from this perspective, they become authors of the collective process and therefore accept responsibility for its outcomes. This approach is distinct from many other "participatory design" processes in which participant's contributions to the design process do not demand a collaborative authoring of its outcomes. Collaborative authoring of outcomes is a value-based design process in which presence as value is key.

### Science and Technology Studies: Actor Network Theory

Science and Technology Studies (STS) studies how science and technological innovation affect society (Hackett et al. 2007). A variety of disciplines and methodologies contribute to STS. The Actor Network Theory, ANT (Latour 2005), is of specific interest for presence as a value for design in a ubiquitous technology and media landscape. ANT argues that causality of what happens next is seldom the result of a direct causal relation. An extensive network, with a variety of cultural, economic, and political dynamics, exists, in which things (material) and concepts (semiotic) contribute to the state of affairs at a certain moment in time. Individuals execute their strive for survival and well-being within such networks. Analyses with ANT shed light on how individuals in the network (consisting of things and concepts) strive for well-being and survival and from this perspective offers insightful presence design analyses mainly focusing on presence as a factor of analysis.

### Political Economy: Poly-centricity

For over 40 years, Elinor Ostrom studied how rural communities in different places in the world become successful and sustainable. Ostrom specifically studied what rules are necessary to create sustainable communities in which individuals have autonomy and in which ecology is balanced. In other words, she studied how communities in which individuals strive for their own well-being and survival can be sustainable with respect for, and in balance with, natural resources. In her research Ostrom concludes there is a limit to how many people can participate in such a community for it to be successful and sustainable. Successful communities meet 8 design

[9]Garrick Jones and Patrick Humphries studied processes of change at the London School of Economics, building upon academic research and business consulting practices.

requirements (Ostrom 1990; Ostrom 2009).[10] When a community becomes too large, it should be split. To this end she developed the notion of poly-centricity, allowing different centers to be autonomous and collaborate at the same time in a network with other communities. Today's network society offers a range of new possibilities for creating such poly-centricity between successful and sustainable communities in which presence functions as key value in Design for Values.

### Participatory Distributed Systems Design: Local Coordination for Global Management

Fundamental to participatory distributed systems design is the notion of local coordination. Every participant moves and acts according to its own interest, steering towards well-being and survival. By accumulating outcomes of all participants steering towards well-being and survival according to certain rules, a participatory system executes its mission (Brazier and Nevejan 2014). For example, a traffic navigation system such as TomTom not only indicates itineraries for car drivers; it also includes real-time data about traffic jams and possible alternative routes to support participants in TomTom's distributed participatory system to adapt their own itineraries for their own well-being. As a result, traffic jams dissolve.

Self-organization and emergence are key to the notion of "local coordination for global management," which is fundamental to complex systems design. Participatory systems design – integrating social, ecological, and technological systems – builds upon principles of complex systems design and specifically adds the value of presence for allowing people to accept responsibility in complex environments (Brazier and Nevejan 2014). In this approach presence as a value for design functions as a design requirement, as a factor of analysis, and as a key value in Design for Values.

## Comparison and Critical Evaluation

Presence design requires the involvement of different scientific and design disciplines. This in itself is a major issue. Connecting psychological, sociological, economic, technological, and cultural designs, such interdisciplinary approaches

[10]Elinor Ostrom's design principles for sustainable communities (stable local pool resource management) are:

1. Clearly defined boundaries (effective exclusion of external unentitled parties)
2. Rules regarding the appropriation and provision of common resources that are adapted to local conditions
3. Collective-choice arrangements that allow most resource appropriators to participate in the decisionmaking process
4. Effective monitoring by monitors who are part of or accountable to the appropriators
5. A scale of graduated sanctions for resource appropriators who violate community rules
6. Mechanisms of conflict resolution that are cheap and of easy access
7. Self-determination of the community recognized by higher-level authorities
8. In the case of larger common-pool resources, organization in the form of multiple layers of nested enterprises, with small local CPRs at the base level

require multilingual capacity between different communities of practice (Kuhn 2000). Even when this multilingual capacity is available, there is no best solution, no ultimate system to be designed. As history shows, we, as human beings, with our ability to strive for well-being and survival continually find new ways to adapt, invent, and move on. Nevertheless, in today's world we are dependent on complex systems that define basic utilities, transport, food and water, finance culture, politics, and more. Presence as a value for design is fundamental to all of these systems, in particular to support emergence as the outcome of the accumulation of many participants' strive for well-being and survival is most often characterized by processes of self-organization and emergence. This in itself is a challenge, as the process of self-organization is, by definition, unpredictable.

The need to integrate our strive for survival and well-being in the design process from the start is implicit in each of the approaches discussed above. Note, however, that we, as human beings, are changing due to our networked societies, with ubiquitous technology in pervasive media landscapes. Such changes pertain not only to our own psychological and physiological being but also to how social structures emerge and function with increasing complexity.

## Three Examples of Presence as Value for (Meta-)design

There are three ways in which values can play a role in a design process: as design requirement, as factor of analysis, and as the value driving a value-sensitive design process (Vermaas et al. 2011; van den Hoven 2005). To shed light on each of these roles, the YUTPA framework is used to analyze and design presence as value for (meta) design (see section "What Does It Mean to Design for Presence?" under "meta design for choices and trade-offs") (see Fig. 1).

Interdisciplinary research has identified 4 dimensions of significance for making choices and trade-offs for the performance of presence. The YUTPA framework, acronym for being with You in Unity of Time, Place and Action, sheds light on specific presence configurations in which a person performs presence with YOU, in the NOW, being HERE, with a specific potential to DO certain things.

Each of the dimensions of relation, time, place, and action is defined by a number of factors, which affect how a person judges the presence configuration in which one finds oneself. As a result specific trust is established, which affects how a person performs presence.

In the dimension of relation, identified factors are role, reputation, engagement, and communion (shared meaning). In the dimension of time, the factors are duration of engagement, integrating rhythm, synchronizing performance, and making moments to signify. In the dimension of place, the factors are body sense, environmental impact, emotional space, and situated agency. In the dimension of action, the factors are tuning, reciprocity, negotiation, and quality of deeds (actions and activities).

The YUTPA framework facilitates discussion about presence configurations. In a YUTPA analysis, appointed levels to each factor are subjective indicators, and

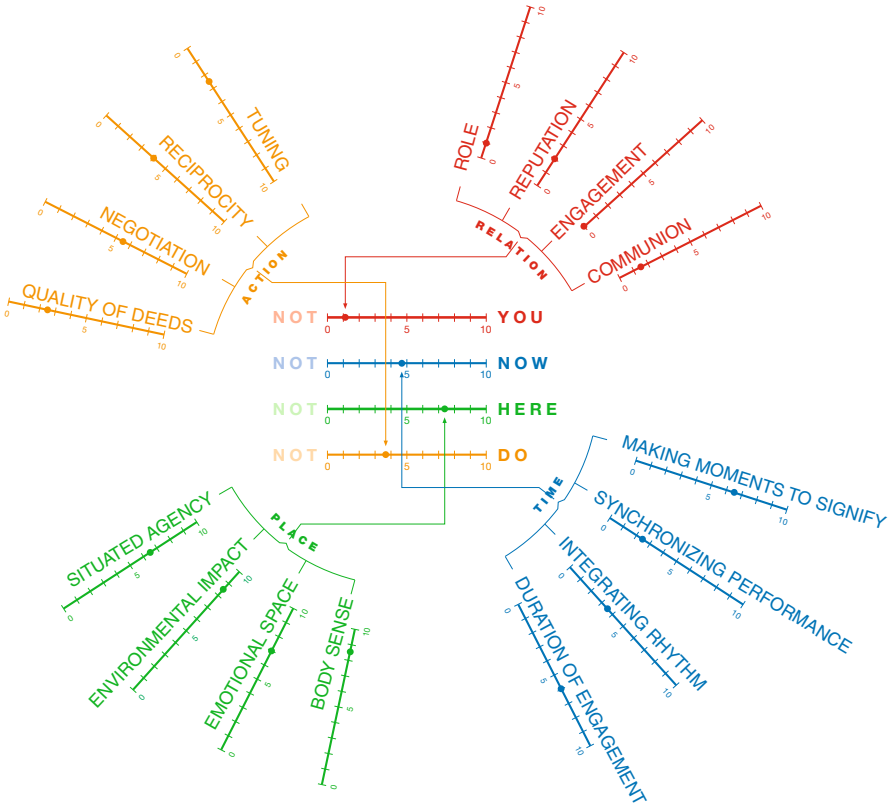



**Fig. 1** The YUTPA framework (design: office of CC, Amsterdam)

not objective calculated outcomes, for facilitating conversation about a specific presence design.

**Example 1: Presence as a Design Requirement – Augmented Reality for Expert Collaboration**

When translating presence into design requirements, an application should facilitate a participant's capacity to steer towards his/her own and others' well-being and survival. A participant's possibilities to act have to be real in the sense that they can be aware of the situation they are in and act upon it. This is one of the great challenges in the design of augmented reality applications in which experts have to collaborate.

## YUTPA Analysis CSI the Hague

The research project CSI The Hague explores the potential of mediated and augmented reality for future crime scene investigation. Using special VR glasses

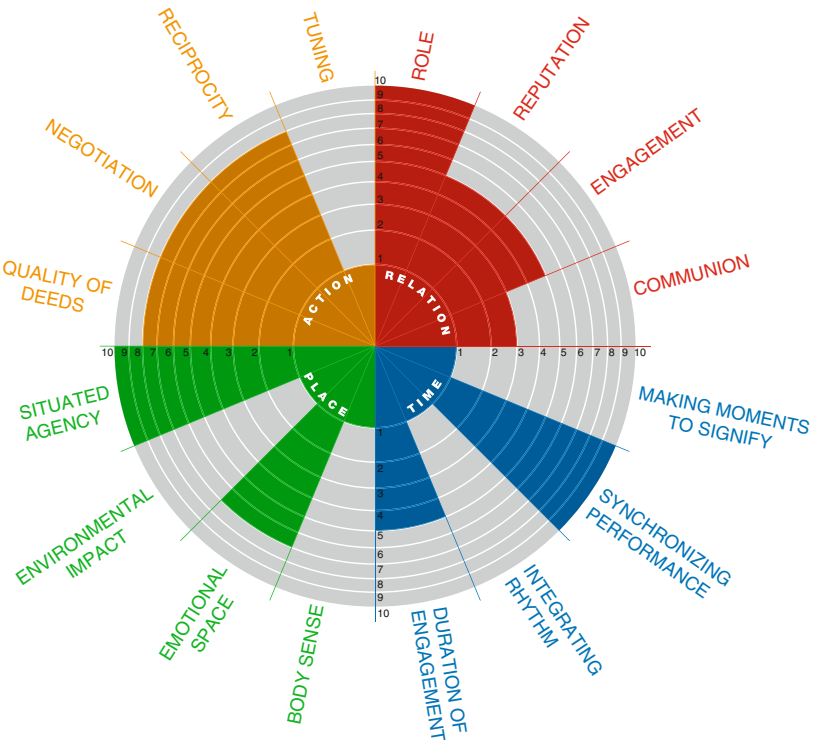



**Fig. 2** This YUTPA analysis shows possible design spaces for CSI The Hague, an augmented reality application for expert collaboration in forensic crime scenes (Design: office of CC, Amsterdam)

through which experts see a real crime scene as well as augmented indicators that colleagues have placed, the application needs to facilitate experts to investigate together (see Fig. 2).

**Relation:** Experts in the Crime Scene Investigation of the CSI The Hague project meet each other in professional roles. This defines their engagement and affects their reputation. Interestingly, the experts in this case need to create a shared meaning, not the same type of shared meaning as the shared meaning we make with family or friends but a shared meaning to contextualize and understand a crime scene investigation that also includes ethical positions in the process.

**Time**: Experts work together for the limited amount of time that is needed to do the investigation. They are trained in their professional roles to synchronize performance. This is not often possible as their rhythms will often not be integrated as they have different professional environments and may even live in different time zones. Because they work online in mediated presence, it is almost impossible to share moments that signify. It is almost impossible to share celebration when successful or share the mourning that comes with atrocity or defeat. On the time dimension, the performance of presence is defined by the lack of trust caused by a low integration of rhythms and not sharing of moments to signify.

**Place:** The body sense and environmental impact between an expert on the crime scene and an expert elsewhere is very different and therefore contribute very little to the collaborative performance of presence. The emotional space experts share depends significantly on the compassion and experience of the remote expert. This emotional space is defined by professional roles, expert knowledge of the task at hand, but also by the gravity of the situation with which both experts have to deal. Situated agency, the fourth factor in the place dimension, is clearly defined as a requirement. The purpose of the application is to give agency to the remote expert to make remote collaboration effective, which, if successful, will significantly contribute to experts trust in the situation.

**Action**: Reciprocity in signs and negotiation of conditions are performed in professional settings and can be executed in remote situations as well. However, these are hindered by the lack of tuning possibilities. No body movements or breath space can be shared. Trust in the expert collaboration may be created by a series of activities for contextualizing the actions experts may exchange at distinct moments in time.

Overall the YUTPA analysis shows that the degree of trust in the expert in augmented collaboration is a challenge. Of the 16 factors that have been identified so far, only 7 contribute significantly to trust affecting choices and trade-offs for presence in the current design. Synchronizing performance, situated agency, reciprocity, negotiation, quality of deeds, and role are requirements on which the design of the system it is based. Depending on the experts that engage with each other, emotional space and communion may contribute to the degree of trust in augmented collaboration in which case the balance flips to more than half of the factors contributing to trusting mediated collaboration. But these are highly individual factors. From a presence design perspective, the system could benefit from the time dimension by enhancing, for example, "integrating rhythm." In the place dimension, situated agency and emotional space could benefit from explicit functionality designed to this purpose. In the action dimension, there are options to improve tuning of presence and quality of deeds. In the relation dimension, a reputation system may contribute to the sense of presence.

**Example 2: Presence as a Factor of Analysis – Facebook**

Presence as a factor of analysis judges the choices that are made in a design process giving agency to participants to steer towards their own and others well-being and survival. Such agency needs to be in balance with attention, intention, and expectation of participants in the to be designed participatory scripts.

## YUTPA Analysis Facebook

In this example a YUTPA analysis is carried out to understand how Facebook's presence design generates trust for its participants (see Fig. 3).

**Relation:** Depending on personal style, all identified factors in the dimension "relation" play a role of significance in networks of friends. Some people use

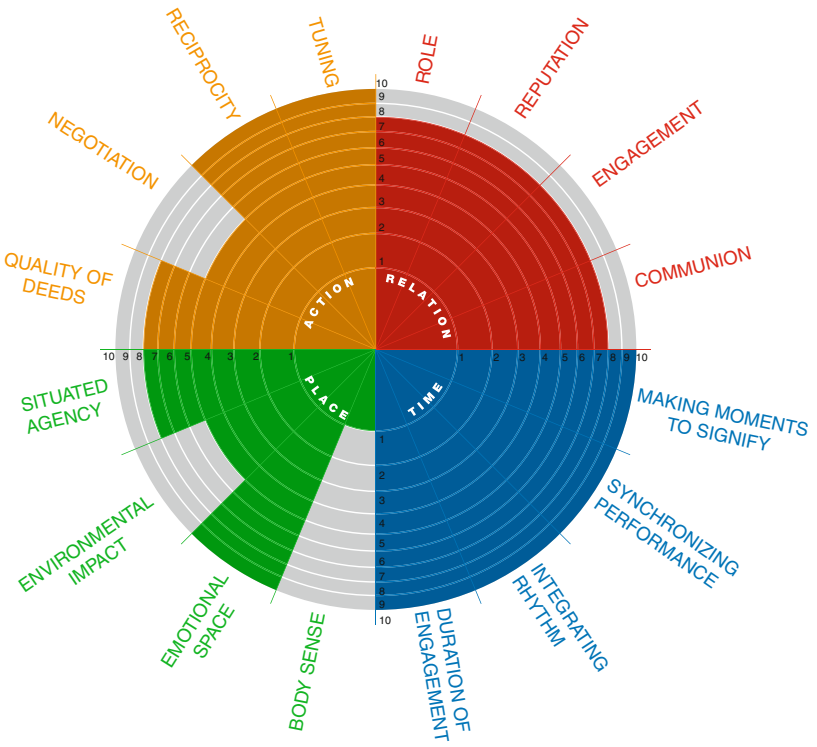



**Fig. 3** This YUTPA analysis focuses on analyzing Facebook from the perspective of presence as value for design (Design: office of CC, Amsterdam)

Facebook mostly professionally for which Facebook scores high with respect to its role. Facebook functions as a reputation system; for example, employers look up possible new applicants to learn more about them. Engagement can be very high, up to the point of addiction. In specific contexts, Facebook is part of creating shared meaning. Family and friends use Facebook to stay in touch. The dimension of relation in Facebook's design contributes significantly to why people trust Facebook.

**Time:** For Facebook the duration of engagement is endless and 24/7. Its design supports integration of posts of friends minute by minute in individuals' own rhythms and activities of the day. Synchronizing with friends, for example, by entering a chat, or issuing likes is instantaneous. Facebook communication is also designed to support significant moments in peoples' lives, for example, when friends celebrate a shared meaning as in protest or a party. The time dimension of Facebook's design generates a high level of trust.

**Place:** Facebook does not directly affect our body sense. It also does not have or create direct environmental impact, but many friends may live in the same environment and therefore Facebook may have environmental impact. The emotional space Facebook offers is immense and elaborate for many. It offers "situated

agency" allowing participants to post, like, and comment on anything they notice. So the dimension of place contributes significantly to the emergence of trust.

**Action:** Depending on personal style and choice, Facebook is able to support intense tuning with others as well as reciprocity between friends. Negotiation is not really one of its features, although some people may invent ways to acquire this functionality within Facebook. Concerning the quality of deeds, it seems that most people use Facebook as part of their daily activities. At some moments in time in specific context, a post may be considered a deed. It is clear though that the dimension of action contributes significantly to the emergence of trust.

From this short analysis, it may be concluded that Facebook generates trust from its participants by its presence design. In this presence design, the dimension of time is crucial. Followed by action and relation, but also the dimension of place contributes significantly. However, a YUTPA analysis does not shed light on political opinions on how Facebook as a company behaves and can be trusted or not. Quite many people do not participate in Facebook because of Facebook's data policy. This policy includes giving details of Facebook users to both to business and intelligence corporations. This YUTPA analysis sheds light on how people make choices for presence and trust but does not incorporate judgments on larger issues of trust as Facebook's behavior as a company, for example.

When judging increase or decrease in presence in a specific design trajectory, arguments need to incorporate social, economic, political, and ecological consequences of the intended participatory scripts for presence. From one perspective it may seem that a participant acquires agency, while, for example, from another perspective actual economic or political circumstances deeply affect the situation in such a way that presence for other participants decreases. The analytical frameworks discussed in section "Meta-design for Choices and Trade-Offs" address the social, political, and economical issues of presence as a value for design. An ANT analysis (Actor Network Theory), for example, identifies relations of Facebook with the world of finance, intelligence, and

**Example 3: Presence in Design for Values – Smart Grid**

Presence as a value in Design for Values positions our "strive for well-being and survival" center stage in all phases of the design process. However, systems necessarily have multiple actors each with their own strive for well-being and survival. Their needs may collide. Where in nature's design, according to Darwin, in the strive for well-being and survival the fittest will survive, in designs for human society, more complex and more balanced presence design is possible. Colliding, interdependent needs of multiple actors need to be taken into account, as the context for design.

For social structures, including businesses, to be sustainable, a balance between individual and collective strive for well-being and survival has to be met. To this end design choices have to be made for modes of participation, modes of communication and decision-making, and modes of influence and authority in the context of network, networked, networking, and network-making powers (Castells 2012). Also this presence design is effectively a

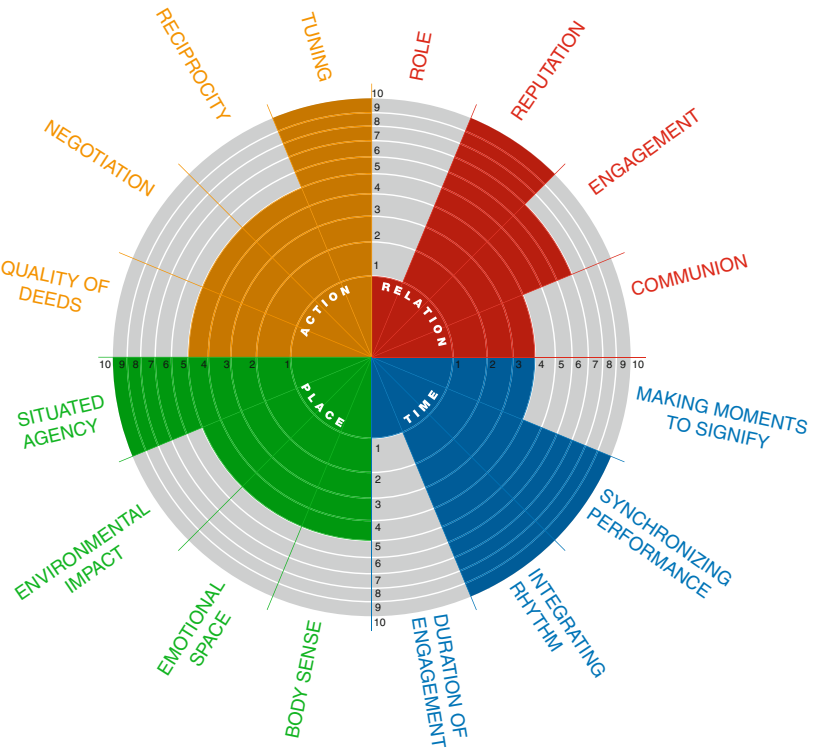



**Fig. 4** This YUTPA analysis unfolds design solution spaces for presence as key value in a value sensitive design process for smart grids (Design: office of CC, Amsterdam)

meta-design in which structures of governance and structures of participation are designed to be amended over time.

The different analytical frameworks, as discussed in section "Meta-design for Choices and Trade-Offs," are all of relevance to Design for Values: presencing, collaborative authoring of outcomes, simulations and emulations for paying tribute to the diverse links in the actor network system, poly-centricity, and distributed systems design. Presence as a value in Design for Values needs to address agency of participants and the potential for trust between participants including the system itself. Being and bearing witness have to be scripted in (Nevejan and Brazier 2014).

## YUTPA Analysis Smart Grid

Currently smart grid technology is developed worldwide. Boulder Colorado, for example, the first smart grid city in the USA, provided two-way connectivity to the city. Citizens can be both consumers and producers of energy and the grid negotiates and divides according to the needs and possibilities of each household (see Fig. 4).

In Western Europe energy is available 24/7. In current energy market "supply" follows "demand." With the expectation that over time, as energy resources and needs change (e.g., with the introduction of electric vehicles), future smart grids have to be designed in such a way that "demand" will follow "supply."

In this example the YUTPA framework is used to identify solution spaces for designing smart grid technology in West European cities.

**Relation**: Currently our role in energy nets is most often as a consumer. As more and more consumers become producers (prosumers), our roles change. Prosumers are more engaged with the energy they use. Critical solution space for designing smart grids is the facilitation of different forms of engagement allowing people and businesses to accept different roles in the production and consumption of energy. Smart reputation design is becoming a factor of significance, for example, related to the contribution prosumers make to societal sustainability. The shared meaning that may emerge as result of being involved with the smart grid offers a solution space for both designed and emerging cultural dynamics that affect communities' energy supply and demand.

**Time:** Personal duration of engagement with the electricity grid is characterized by moments of intense use and periods of nonuse. However, electrical current is available 24/7. Because of its "continual availability," there is no need to integrate our personal rhythms with nature (day and night, cold and warm) or with our neighbors for more efficient energy use.

When energy is less abundant, two factors in the time dimension offer solution spaces for design. Smart integration of rhythms between people, communities, businesses, and geographical regions offers new opportunities for efficient energy production and energy use.

Secondly, synchronization of performance between supply and demand creates a solution space that can be explored in which new developments in ICT (mobile networks, sensor technology, and agent-based platforms) can significantly contribute. Concerning "moments to signify," the current electricity net is seen as a utility that has to function 24/7. As consumers become prosumers, and actively both consume and produce energy, new moments to signify may emerge. Designing new moments for signifying the way individuals and communities handle their energy use may contribute to a new culture of energy and play a role in strategies for change.

**Place:** The experience of energy is bound to the place where the body resides. Body sense and environmental impact are fundamental to energy use. Energy keeps us warm, allows us to read at night, and makes ICT function. However, body sense and environmental impact are more a given than a solution space for design. Emotional space is not directly influenced, although the effect of no energy directly affects personal and relational spheres as indicated by the urban myth that a baby boom takes place 9 months after an energy black out. Situated agency is defined, for most, by turning on a switch anytime during the day and by paying (automatically) a bill once a month. The feedback to our "energy actions" is immediate; a light turns on. The expense of our "energy actions" is very remote; the bill comes weeks later. In a smart grid situation, local production of energy affects actions in day-to-day

life, "demand" follows "supply," and feedback is experienced in the here and now. Surplus energy is traded locally, regionally, or even globally, but all benefits are awarded locally. Design solution spaces in the place dimension for smart grids are mostly defined by the bandwidth for situated agency.

**Action:** The quality of deeds concerning energy are very diverse: cutting a tree for a fire, mining coal under the ground, executing operations in a nuclear plant, and placing a solar panel on the roof or turning the windmill towards the wind all have deep impact on day-to-day lives. Tuning human behavior and the production of energy is one of the possibilities that smart grid technology provides. Developments in ICT (sensors, Internet of things, big data, agent technology) support personal and local aggregation of data on the basis of which energy use can be tuned/aligned with human behavior.

Where today negotiation of energy resource provisioning (and prices) is mainly the domain of the electricity companies, communities of consumers are emerging in which energy production and consumption is negotiated. In such energy communities, reciprocity in exchanging energy resources between participants directly affects the lives of the participants. Each of these factors (tuning, negotiation, reciprocity, and quality of deeds) offers solution spaces for design of smart grids.

This short YUTPA analysis shows there are many design spaces for smart grid technologies in which individual presence and collective strive for survival and well-being are intertwined and where this interdependency can be fruitful.

Faced with different social and ecological crises, understanding the design space for presence is fundamental for social structures of the future. It should drive innovative solutions for next-generation infrastructures. Our participation in complex distributed architectures and infrastructures requires the taking of responsibility. Having the possibility to enact our own presence, to execute our own agency is fundamental to infrastructures, architectures, and governance structures that rely on us to take responsibility and accept accountability. Presence design as value-sensitive design emphasizes participation as a way to manifest presence of participants involved.

## Open Issues and Further Work

Little is known about the embodiment of virtual and mediated experiences. Looking at data of users, it seems that millions of people engage daily in network activities. How such activities affect us human beings is unclear. Effects on human psychology, on how communities and societies function, on how markets adapt, and many more questions are open issues and subject of further research. How does network reality influences the mind maps we make? How does networked reality become embodied? How does network reality affect our feelings and emotions and are emotions and feelings also fundamental to steering in network realities, or are there other drivers in the online world? The relation between performance of presence and imagination needs to be explored much deeper for being able to answer question like this.

Network reality is part of our daily negotiation for performance of presence, but most psychological and sociological theories are based on a world in which network reality does not play a role. It is unclear whether psychological and sociological mechanisms can be transposed to network reality. In mass communication, in media studies and in net critique, these issues are being explored; a new paradigm for analysis and design is emerging but is not yet clearly defined.

A confusing issue is that we, as people, negotiate performance of presence based on how we trust a situation with which we are confronted. Trust may not be granted for appropriate reasons and performance of presence may not be beneficial in the end. As in the Facebook example above, it is easy to trust Facebook because its presence design supports us to significantly steer in each of the 4 dimensions of time, place, action, and relation. However, this trust may be misleading. Power relations in the network society are often opaque, but not less relevant. It is an open issue how distributed transparency can be designed. Also it is unclear how we, as individuals, are positioned in personal-global dynamics. Whistle-blowers like Edward Snowdon and Julian Assange show how technology in the hands of a few controls many. They reach a large audience via the media, but little political action happens as a result. These open issues have great societal impact and further research is timely.

## Conclusions

This chapter focuses on the design of presence in merging realities as approached in the social and design sciences. Presence is a fuzzy concept. Many methodologies implicitly include or exclude presence as value for design.

Current presence research focuses on creating the sense of presence in being there, but it most often does not address larger issues of societal impact of presence design. In our day-to-day lives in social networks and pervasive ubiquitous technologies upon which fundamental processes of life depend in network societies, on- and offline realities merge. The being here and the being there are one in human experience.

Presence design is not design for specific behavior for presence; it is meta-design; it is designing for choice and trade-offs between choices. It is design for experience in which current and historical contexts are taken into account together with actual perceptions and understanding. Both scientific design and artistic research contribute to presence design.

Despite all of the current models of thinking, the current speed and scale of technological innovation is changing our lives profoundly. It is as if we are part of a global experiment in which dynamics of information, communication, and trans-action, all fundamental to society, are changing, dynamics that have existed for over a thousand of years of building up experience and social structures, markets, and structures of governance to be able to live together. Today systems of law and systems of value exchange are all under pressure. We, as human beings, are changing as result of the global network society.

The strive for well-being and survival is deep in our DNA and will keep on defining what will happen next. By incorporating presence as a value for design, and configuring design processes accordingly, "old" human experience will have a chance to resonate and inform future generations to come for designing and creating a social, technological, and ecological environment worth living in.

## Cross-References

► Design for the Value of Trust

## References

Azuma R (1997) A survey of augmented reality. Presence 6 4:355–385
van Baren J, IJsselsteijn W (2004) Measuring presence: a guide to current measurement approaches. Report for the OmniPres project, funded by the European Community under the Information Society Technologies Programme
Baudrillard J (1983) Fatal strategies (trans: Beitchman P, Niesluchowski WGJ). In: Semiotext(e), a history of the present
Baxi U (1999) Voices of suffering, fragmented universality, and the future of human rights. In: Weston BH, Marks SP (eds) The future of international human rights. Transnational Publishers, New York, pp 101–156
Benjamin W (1936) Het kunstwerk in het tijdperk van zijn technische reproduceerbaarheid (trans: Hoeks H). SUN, Nijmegen 1985. Translated from: Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit SuhrkampVerlag, Frankfurt am Main 1974
Biggs M, Karlsson H (eds) (2011) The routledge companion to research in the arts. Routledge, London
Borgdorff H (2012) Conflict of the faculties, perspectives on artistic research and academia. Leiden University Press, Amsterdam
Boyd D (2014) It's complicated: the social lives of networked teens. Yale University Press, New Haven
Brazier F (2011) Inaugural speech Delft University of Technology, Oct 2012, www.particpatorysystems.org
Brazier F, Nevejan C (2014) Framework for participatory systems design. Paper presented at CESUN 2014, Stevens Institute for Technology, New Jersey
Butler J (1993) Bodies that matter, on discursive limits of "sex". Routledge, New York
Castells M (1996) The rise of the networked society. Blackwell Publishing, Oxford
Castells M (2001) The internet galaxy, reflections on the internet, business and society. Oxford University Press, Oxford
Castells M (2009) Communication power. Oxford University Press, Oxford/New York
Castells M (2012) Networks of outrage and hope. Social movements in the internet age. Polity Press, Cambridge
Damasio A (2000) The feeling of what happens. Body, emotion and the making of consciousness. Vintage, Random House, London
Damasio A (2004) Looking for Spinoza, joy, sorrow and the feeling brain. Vintage, Random House, London
Dewey J (1934) Art as experience. Penguin, New York. Perigee Books (1980)
Fischer G (2013) Meta-design: empowering all stakeholder as co-designers. In: Luckin R, Goodyear P, Grabowski B, Puntambeker S, Underwood J, Winters N (eds) Handbook on design in educational computing. Routledge, London, pp 135–145

Floridi L (2005) The philosophy of presence: from epistemic failure to successful observation. In: Presence: teleoperators & virtual environments- special section: legal, ethical and policy issues associated with virtual environments and computer mediated reality, vol 15(6). MIT Press, Cambridge, pp 656–667
Giddens A (1984) The constitution of society, outline of the theory of structuration. Polity Press, Cambridge
Gill SP, Nevejan C (2012) Introduction to special issue witnessed presence. AI Soc J Knowl Cult Commun 27(Special issue Witnessed Presence):6
Goffman E (1963) Behavior in public places; notes on the social organization of gatherings. The Free Press, New York
Gullstrom C (2010) Presence design: mediated spaces extending architecture. PhD dissertation, KTH, Stockholm
Habermas J (1983) Moral consciousness and communicative action (trans: Lenhardt C, Nicholsen SW) (1990). Introduction by Thomas A. McCarthy. The MIT Press, Cambridge
Hackett E, Amsterdamska O, Lynch ME, Wajcman J (eds) (2007) The handbook of science and technology studies, 3rd edn. MIT Press, Cambridge
Hamelink CJ (2000) The ethics of cyberspace. Sage Publications, Thousand Oaks/London/New Delhi. Original title: Digitaal Fatsoen. Uitgeverij Boom, Amsterdam 1999
Hendrix C, Barfield W, Bystrom KE (1999) A conceptual model of the sense of presence in virtual environments. Presence Teleoper Virtual Environ 8(2):241–244
van den Hoven MJ (2005) Design for values and values for design. Information age +. J Aust Comput Soc 7(2):4–7
Humphries P, Jones G (2006) The evolution of group decision support systems to enable collaborative authoring of outcomes. In: World futures: the journal of general evolution, vol 62. Routledge, Oxford, pp 171–192
IJsselsteijn WA (2004) Presence in depth. PhD dissertation, Technische Universiteit Eindhoven
IJsselsteijn W, Giuseppe R (2003) Being there, the experience of presence in mediated environments. In: Giuseppe R, Fabrizio D, Wijnand I (eds) Being there: concepts, effects and measurement of user experience in synthetic environments. IOS Press, Amsterdam, pp 3–16. Retrieved from http://www.ijsselsteijn.nl. Accessed 30 Nov 2006
Ilavarasan PV (2008) Software work in India: a labour process view. In: Upadhya C, Vasavi AR (eds) An outpost of the global economy, work and workers in India's information technology industry. Routledge, New Delhi
Kuhn TS (2000) In: Conant J, Haugeland J (eds) The road since structure, philosophical essays, 1970–1993, with an autobiographical interview. The University of Chicago Press, Chicago
Latour B (2005) On the difficulty of being ANT: an interlude in the form of a dialogue. In: Reassembling the social. Oxford University Press, Oxford
Lombard M, Jones MT (2007) Identifying the (Tele)Presence literature. PsychNology J 5(2):197–206
Lovink G (2012) Networks without a cause: a critique of social media. Polity Press, Cambridge/Malden
Lunenfeld P (2003) The design cluster, preface. In: Laurel B (ed) Design research: methods and perspectives. MIT Press, Cambridge/London, pp 10–15
Mann S, Barfield W (2003) Introduction to mediated reality. Int J Hum-Comput Interact 15:205–208
McLuhan M (1964) Understanding media: the extensions of man. McGraw-Hill, New York
Mensvoort K, Grievink H-J (2012) Next nature. Actar, Barcelona
van der Meulen S (2012) Witnessed presence in the work of Pierre Huyghe. AI Soc J Knowl Cult Commun 27(Special issue on Witnessed Presence):1
Negt O, Kluge A (1972) Öffentlichkeit und Erfahrung: Zur Organisationsanalyse von bürgerlicher und proletarischer Öffentlichkeit. Suhrkamp, Frankfurt am Main
Nevejan C (2007) Presence and the design of trust. PhD dissertation, University of Amsterdam
Nevejan C (2009) Witnessed presence and the YUTPA framework. PsychNology J Ethics Presence Soc Presence Technol 7(1). www.psychnology.org

Nevejan C (ed) (2012) Witnessing you, on trust and truth in a networked world, Participatory systems initiative. Delft University of Technology, Delft
Nevejan C, Brazier F (2010) Witnessed presence in merging realities in healthcare environments. In: Advanced computational intelligence paradigms in healthcare 5. Studies in computational intelligence, 2011, vol 326/2011. Springer, pp 201–227
Nevejan C, Brazier F (2011) Granularity in reciprocity. AI Soc J Knowl Cult Commun 27:1
Nevejan C, Gill SP (2012) Special issue on witnessed presence 2012. AI Soc J Knowl Cult Commun 27:1
Nowak KL, Biocca F (2003) The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. In: Presence: teleoperators and virtual environments, vol 12(5). MIT Press, pp 481–494
Oliver K (2001) Witnessing, beyond recognition. University of Minnesota Press, Mineapolis/London
Ostrom E (1990) Governing the commons: the evolution of institutions for collective action. Cambridge University Press, Cambridge
Ostrom E (2009) A general framework for analyzing sustainability of social-ecological systems. Science 325:419–422
Poelman R, Akman O, Lukosch S, Jonker P (2012) As if being there: mediated reality for crime scene investigation. In: CSCW '12: Proceedings of the ACM 2012 conference on computer supported cooperative work. ACM, New York, pp 1267–1276
Postman N (1985) Amusing ourselves to death: public discourse in the age of show business. Penguin, New York
Riva G (2008) From virtual body to real body: virtual reality as embodied technology. J Cyber Ther Rehabil 1(1):7–22
Riva G, Waterworth JA, Waterworth EL (2004) The layers of presence: a bio–cultural approach to understanding presence in natural and mediated environments. CyberPsychol Behav 7(4):402–416
Schwab M, Borgdorff H (eds) (2014) The exposition of artistic research: publishing art in academia. Leiden University Press, Leiden
Senge P, Scharmer CO, Jaworski J, Flowers BS (2004) Presence, exploring profound change in people, organizations and society. Published by arrangement with Society for Organizational Learning
Short JA, Williams E, Christie B (1976) The social psychology of telecommunications. Wiley, New York
Slater M (2014) Grand challenges in virtual environments. In: Frontiers in Robotics an AI. http://journal.frontiersin.org/Journal/10.3389/frobt.2014.00003/full. Accessed 23 Aug 2014
Spagnoli A, Gamberini L (2004) The sense of being 'There': a model for the space of presence. In: Raya MA, Solaz BR (eds) Paper presented at the seventh annual international workshop presence 2004, October 13–15, in Valencia, Spain. In printed collection of papers, Editorial de la Universidad Politecnica de Valencia, Valencia, pp 48–53
Spanlang B, Fröhlich T, Descalzo FF, Antley A, Slater M (2007) The making of a presence experiment: responses to virtual fire. In PRESENCE 2007 – the 10th annual international workshop on presence, Barcelona, 2007
Upadhya C (2008) Management of culture and management through culture in the Indian outsourcing industry. In: Carol U, Vasavi AR (eds) In an outpost of the global economy, work and workers in India's information technology industry. Routledge, New Delhi
Virilio P (1989) Het Horizon negatief (trans: Mulder A, Riemens P). Amsterdam: Uitgeverij Duizend & Een. Original title: L'Horizon negatif: essai de dromoscopie Galilee 1984, Paris
Weston BH, Marks SP (eds) (1999) The future of international human rights. Transnational Publishers, New York
Whiteside AL, Garret Dikkers A (2012) Using the social presence model to maximize interactions in inline environments. In: Amant KS, Kelsey S (eds) Computer-mediated communication

across cultures: international interactions in online environments. IGI Global, pp 395–413. http://onlinelearningconsortium.org. Accessed 21 Sept 2014

Wiederhold BK (2006) The potential for virtual reality to improve health care. The Virtual Reality Medical Center, San Diego

Wyatt S (2004) Danger! Metaphors at work in economics, geophysiology and the internet. Sci Technol Hum Values 29(2):242–261

Zijlmans CJM (2013) Laboratory on the move in retrospect. In: Thissen J, Zwijnenberg R, Zijlmans CJM (eds) Contemporary culture. New directions in art and humanities research. Amsterdam University Press, Amsterdam, pp 175–186

# Design for the Value of Privacy

Martijn Warnier, Francien Dechesne, and Frances Brazier

## Contents



**Abstract**

In a time where more and more information about people is collected, especially in the digital domain, the right to be left alone and to be free of surveillance, i.e., privacy, is no longer as self-evident as it once was. Therefore, it is important that new systems are designed with privacy in mind. This chapter explores the notion of privacy and how to design "privacy-preserving" systems: systems that are designed with privacy for the end users in mind. Several design approaches that address this issue, such as "Privacy by Design," "Value Sensitive Design," and "Privacy Enhancing Technologies," are discussed. Examples of privacy-preserving (and breaking) systems, ranging from smart meters to electronic health records, are used to illustrate the main difficulties of designing such systems.

M. Warnier (✉) • F. Dechesne • F. Brazier
Delft University of Technology, Delft, The Netherlands
e-mail: m.e.warnier@tudelft.nl; f.dechesne@tudelft.nl; f.m.brazier@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_17

**Keywords**
Privacy • Design • Value Sensitive Design • Smart grid

## Introduction

Throughout history only a privileged few enjoyed the privacy that in recent times has become more commonplace: the right to be left alone and not be under surveillance, both from peers as well as governments. In the last decades, this has changed again with the rise of the Internet. What began as a means to freely and anonymously communicate with others around the world has become an instrument for violating the privacy of individuals at a scale hitherto not thought to be possible. Developments in information technology, such as increasing computing power, storage, and communication, have led to many benefits for people, but individual privacy has come under threat. All kinds of data, ranging from marketing information (buyer profiling) to medical data, are collected, linked, and processed both by companies and governments. The increasing connectedness of stored data makes it possible to link more data to individuals, thereby stretching what counts as "personal data."

The right to privacy (Warren and Brandeis 1890) is a universal human right (Movius and Krup 2009). It entails both freedom of intrusion or "the right to be left alone" and control of information about oneself. The (computer) systems that do the collection and processing of data should therefore be designed with care for privacy. Designing such systems that preserve privacy is a difficult task (if possible at all), in particular when the system is centered on the processing of privacy-sensitive data (such as medical information). Fortunately, there is a long history of security principles and legislative work that can be used as a starting point for designing such systems for privacy.

The easiest way to design a privacy-preserving system is to not collect, store, or process any personal data. However, in practice many computerized systems need to process some personal data. For a large subset of these systems, there is no direct explicit need to use personal data, i.e., there are no *functional* requirements to the system to collect, store, and process personal data. For example, public transportation systems often use computerized tokens, such as the Oyster system of the London underground or the Dutch OV-chip card, which users have to use to gain access to the public transport system. It can be useful, for example, for future planning or optimization purposes, for such systems to collect data about the number of travelers per train. But there is no reason – except for a commercial one – to store the entire travel history for each individual user (as the Dutch system does). Systems that collect personal data for commercial reasons usually do this to be able to provide personalized (targeted) advertisements or to sell the collected data to other interested parties such as advertisers or insurance companies. Large data processors such as Google and Facebook (but also many less known ones) specialize in this: they are *designed to break privacy* – in particular, users lose control about their own information. They are given an incentive to "trade away"

(part of) their privacy (control over personal data) in exchange for small monetary discounts (Groupon) or specific services (Google, Facebook).

Designing for privacy is not limited to only computer systems; some systems such as RFID tags (Juels 2006), smart phones, and the Internet of things (Atzori et al. 2010) combine physical devices with computer back ends which leads to all kinds of complications in (privacy-preserving) systems design. Other examples such as DNA sequencing have no direct relation with computer systems, but clearly have a privacy impact (and the, privacy sensitive, results of these techniques are often stored in computer systems). For all these systems, it is important to design rules and guidelines that enforce the privacy of the users (or subjects) of the system. Section "What Does It Mean to Design for Privacy?" discusses such a system and its privacy implications, at the border of computer and physical system: the smart grid.

This chapter explores the notion of privacy and how to design "privacy-preserving" systems: systems that are designed with privacy in mind and systems that can be used to circumvent the large data collectors such as Google and Facebook. Examples of privacy-preserving (and breaking) systems, ranging from smart meters to electronic health records, are used to illustrate the main difficulties of designing such systems.

## Privacy

There is no commonly accepted definition of the concept "privacy." Perhaps this is not surprising since the concept is widely studied in such diverse fields as philosophy, law, social sciences, and computer sciences. This section provides a definition of "privacy" that should be acceptable to most. More esoteric – less accepted – notions related to privacy are also discussed.

### Existing Relevant Definitions, Conceptualizations, and Specifications of Privacy

The concept of privacy can be defined in numerous ways and from various perspectives. This chapter discusses the concept of privacy from a philosophical (ontological, ethical) and a legal perspective.

From an ontological perspective, it is clear that "privacy" is a social and indeed a cultural (Zakaria et al. 2003; Liu et al. 2004) construct: without other people, the concept of privacy is meaningless. Privacy is also a right – indeed a fundamental human right (Movius and Krup 2009) – and as such it can be claimed and enforced through legal means. The following three aspects aim to capture the main points associated with the concept of "privacy."

1. Freedom from intrusion, the right to be left alone
2. Control of information about oneself
3. Freedom from surveillance, the right to not be tracked, followed, or watched (in one's own private space)

The first of the above aspects is identical to what Isaiah Berlin called "negative liberty":

> Liberty in the negative sense involves an answer to the question: 'What is the area within which the subject — a person or group of persons — is or should be left to do or be what he is able to do or be, without interference by other persons. (Berlin 1958)

Negative liberty, and thus also privacy, strives for freedom from external constraints. It deals with relations between people (social!). Individuals typically want to be left alone by larger groups such as organizations and states. In contrast, "positive liberty" is defined as freedom from "internal constraints" such as social and culture structures. This is sometimes also explained as the freedom to express oneself as one wants (self-mastery). Privacy can be seen as a necessary precondition for self-expression and thus for positive liberty, as argued by van den Hoven en Vermaas (2007). In this view, privacy is seen as *respect for moral autonomy*, the autonomy to write one's own history and identify with our own moral choices without "critical gaze, interference of others" (van den Hoven en Vermaas 2007).

The second and third aspect of privacy, as defined above, are more closely linked to legal notions of privacy. These deal with the control and storing/capturing of information about individuals. Regulations, guidelines, and laws such as the EU Data Directive (Birnhack 2008; EU Directive 1995) and the United States Federal Trade Commission's Fair Information Practice Principles (Annecharico 2002) try to capture these two aspects in a number of rules, including (i) transparency (How is data stored/processed?), (ii) purpose (Why is data stored/processed?), (iii) proportionality (Is this necessary for this goal?), (iv) access (What do they know about me, can I change it?), and (v) transfer (Who else has access?).

Different countries have different ways of implementing these principles in laws and regulations. For example, the EU has a very strict privacy regulation (the EU Data Protection Directive 1995), that is, enforced "top-down" for all organizations and citizens in the whole European Union. In contrast, regulations in the United States are typically more sector specific such as HIPAA (1996) for the healthcare sector and the Gramm–Leach–Bliley Act (Janger and Schwartz 2001) for the financial sector. Moreover, the United States favor self-regulation, for example, the PCI-DSS (2009) that is used in the credit card sector. Also note that such laws and regulations are not static (legal) objects, and they are continuously being updated, for example, a new version of the EU Data Directive (EU Proposal 2012) has been proposed (also see the next section).

The right to privacy is, at least to a certain degree, relative. One can have a reasonable expectation of privacy in one's own home (see the third aspect above), but not necessarily in public spaces. People that live in the public eye – royalties and celebrities – also have less expectations of privacy in the current, media-centered society. Note that this makes privacy a context-dependent notion.

For privacy, the context of use and control of information is captured in notions such as "spheres of justice" or "spheres of access" (van den Hoven 1999; Nagenborg 2009) and "contextual integrity," as used by Ackerman et al. (2001) and Nissenbaum (2010). What all these notions have in common is that they

interpret privacy in a local context. The meaning and value of information has a local (possibly cultural) aspect which should be taken into account when analyzing privacy. Nissenbaum in particular understands privacy in terms of context-relative information norms, and distinguishes norms of appropriateness, and norms of distribution. She defines contexts as "structured social settings, characterized by canonical activities, roles, relationships, power structures, norms (or rules), and internal values (goals, ends, purposes)" (Nissenbaum 2010, pp. 132–134). The role of context as it relates to privacy is particularly important when it comes to the use of "privacy-preserving technologies (PETs)," discussed further in section "What Does It Mean to Design for Privacy?."

The above definition of privacy is, by intension, rather broad. Others have a slightly more narrow definition. For example, the definition given by the Value Sensitive Design (Friedman and Kahn 2002) approach is: [Privacy] "refers to a claim, an entitlement, or a right of an individual to determine what information about himself or herself can be communicated to others" (Schoeman 1984). Note that this definition only captures the second aspect of privacy.

One last aspect related to privacy is that of incentives: large-scale socio-technical systems have many stakeholders, each with their own incentives – also with respect to privacy. End users of data processing systems sometimes will be given an incentive to give up some (control over) of their privacy in exchange for a monetary discount or service. Many large data processors (Facebook, Google, Groupon) base their business model on this "privacy information for something else" exchange. This issue is also discussed in more detail below.

## Main Issues of Contention/Controversy

While the right to privacy is considered to be a fundamental human right (Movius and Krup 2009), this right is certainly not absolute. As already mentioned, the right to privacy is less relevant in public spaces or for public figures. It is not clear how far this "lack of the right to privacy" can be stretched: courts will penalize journalists and others that have gone too far in this respect. These lines are dynamic and are continuously redefined as society changes.

Also, since (the right to) privacy is considered to be a legal construct, governments can implement (and have implemented) various laws and regulations that are in conflict with the right to privacy. For example, phone taps or other surveillance techniques can be legal in certain jurisdictions as long as specific rules are followed or a court has allowed the phone tap. Governments, the proverbial "big brothers," typically do not respect their own privacy regulations. Depending on the type of government, ranging from open societies to dictatorships, more restrictive and anti-privacy measures are in place. Of course, in practice (at least in open societies) regulations will only allow governments to monitor its citizens as far as deemed "reasonable and necessary" for law and order purposes. Interpreting what is "reasonable and necessary" monitoring (and other anti-privacy measures) is ultimately decided by the courts.

In cases where privacy regulations are clearly in place, it can still be difficult for citizens to also claim this right. Companies and other organizations are obliged by law (at least in the EU) to inform citizens of all the data they have about them, if so requested. However, in practice most companies do not reply to such information requests or give very limited and incomplete information at best (Jones and Soltren 2005; Phelps et al. 2000). So while citizens have the right to control information about them, this right is not actively enforced. A court order can change this, but this is a relatively big hurdle, especially if one considers that hundreds of organizations store (and share!) personal data about citizens.

The newly proposed EU Data Directive (EU Proposal 2012) tries to remedy this situation by including, among others, regulations that enforce disclosure of information about data breaches within 24 h after the data breach became known and regulations that enforce the "right to be forgotten." The latter should, for example, enable citizens to force companies (Facebook, Google, etc.) to remove all stored data they have about themselves. However, even if this proposal becomes EU law, there are still a number of problems (Rosen 2010): first of all, the regulation is again difficult to enforce. Companies can claim that they removed all personal data about an individual, but there is no realistic way that this can be verified. Indeed, removing all backup copies (of to-be-removed data) can be a difficult problem in itself. Moreover, there is also the risk that this right can be used to "rewrite history": it is only a short step from removing information from Facebook to removing information from Wikipedia.[1] Note that the context is again important here.

Another related aspect of privacy deals with the perception that people have of, potentially privacy invading, technologies and their use and in how far "privacy" addresses their moral worries. Often people are not so much concerned with "privacy" in the sense of being left alone but want to be protected from harm or unfair treatment. Van der Hoven and Vermaas (2007) identified four reasons that often ground calls for privacy: prevention of information-based harm, prevention of informational inequality, prevention of informational injustice, and respect for moral autonomy. In this view, people are not primarily concerned about their privacy when they use a system such as Facebook but rather are concerned about what is done with their personal data, which could harm or discriminate them.

A final point of contention is what actually counts as personal data. In the EU Data Protection Directive, personal data is defined as:

> any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity. (art. 2 a, EU Data Protection Directive 1995)

[1]As an example, consider the case of Wolfgang Werle. Werle has been convicted for murder in Germany. He used German privacy laws to sue Wikipedia to get this information removed from his German Wikipedia page. After winning the case, Werle's German Wikipedia page no longer exists, but the information is still accessible from, among others, the English and Dutch Wikipedia pages.

This definition is intentionally left very broad, but it is not clear if it also holds for aggregate data. Such data can, in principal, no longer be used to identify a person. Yet it can still be perceived as an invasion of one's privacy when aggregate data is, for example, used to refuse a mobile phone contract based on your address and aggregate data about the credit worthiness of your postal area. Another related issue is that sometimes, aggregate data can be decomposed into personal data (de-aggregation). This is similar to the problem of (de-)anonymizing discussed below.

## What Does It Mean to Design for Privacy?

Privacy and privacy-preserving technologies have been studied for decades in the field of computer science (Feistel 1973). This section discusses some of the main principles behind these technologies, and how to design new ones.

### Existing Approaches and Tools

The field of computer security has many adages such as "security is not an add-on feature" that stress that security has to be "designed-in" from the start. The same holds true for privacy. In essence there are three different ways to design a (computer) system that respects the user's privacy:

1. Never store any personal information
2. Follow very strict (privacy) rules when storing and processing personal data
3. Only store and process anonymized personal data

The first of these rules obviously works and is by far the surest way to design systems that are "privacy-proof." Unfortunately, it is not always desirable or indeed possible to not store or process any personal data. Many organizations and companies need to store some customer data, ranging from banks to tax offices and hospitals.

For systems that need to handle personal information, the second rule above applies. There are several rules, guidelines, or best practices for designing privacy-preserving systems. Most of these are very general and can be traced back to the principles that are formed by the EU Data Directive: transparency, it should be clear what information is stored; purpose, it should be clear for what purpose the personal data is stored; proportionality, only relevant data should be stored; access, the user should know what personal data about them is stored and they should be able to change errors; and transfer, personal data should only be transferred with explicit permission of the user and the user should be able to request a transfer of personal data. Others, such as the PCI-DSS (PCI 2009), for example, give very detailed guidelines for privacy and security sensitive systems design for a limited domain (in this case that of the credit card industry and its partners such as retailers and banks). Another source of best practices and (security) guidelines for the design of

privacy-preserving systems is provided by various ISO Standards (Hone and Eloff 2002). In addition, the "Privacy by Design" approach as advocated by Cavoukian (2009) and others also provides high-level guidelines in the form of seven principles for designing privacy-preserving systems. Example principles are "Privacy as the Default Setting" and "End-to-End Security – Full Lifecycle Protection" along with the principles (transparency, proportionality) discussed before. The principles of the Privacy by Design approach take as central notion the idea that "data protection needs to be viewed in proactive rather than reactive terms, making privacy by design preventive and not simply remedial" (Cavoukian 2010). Privacy by design also advocates that data protection should be central in all phases of product life cycles, from initial design to operational use and disposal. The Value Sensitive Design approach to privacy (Friedman et al. 2006) proposes similar rules, such as informed consent, i.e., give users the option on what information is stored (or not), and transparency, i.e., tell users which information is stored about them.

Furthermore, the principles or rules that are formed by the EU Data Directive are themselves technologically neutral. They do not enforce any specific technological solutions. As such they can also be considered as (high-level) "design principles." Systems that are designed with these rules and guidelines in mind should thus – in principle – be in compliance with EU privacy laws and (up until a point) respect the privacy of its users. Note that there is a difference between the design and the implementation of a (computer) system. During the implementation phase, software bugs are introduced, some of which can be (mis)used to break the system and extract private information. How to implement bug-free computer systems[2] remains an open research question (Hoare 2003). This issue is further discussed in the next section.

The third rule ("only store and process anonymized personal data") above consists of two different approaches: (i) anonymizing tools such as Tor (Dingledine et al. 2004) and Freenet (Clarke et al. 2001) and (ii) more general, non-technological ways for anonymizing existing data. For example, patient names can be removed from medical data for research, and age information can be reduced to intervals: the age 35 is then represented as falling in the range 30–40. The idea behind this is that a record can no longer be linked to an individual, while the relevant parts of the data can still be used for scientific or other purposes.

Software tools, such as Tor and Freenet, allow users to anonymously browse the web (with Tor) or anonymously share content (Freenet). Such software tools are usually, somewhat misleadingly, called privacy enhancing technologies (PETs). They employ a number of cryptographic techniques and security protocols in order to ensure their goal of anonymous communication. Technically, both systems use the property that numerous users use the system at the same time. In Tor, messages are encrypted and routed along a number of different computers, thereby obscuring the original sender of the message (and thus providing anonymity). Similarly, in Freenet content is stored – in encrypted form – among all users of the system.

[2]Or indeed, how to verify the absence of bugs in computer systems.

Since users themselves do not have the necessary decryption keys, they do not know what kind of content is stored, by the system, on their own computer. This provides plausible deniability and privacy. The system can at any time retrieve the encrypted content and send it to different Freenet users.

A relatively new, but promising, technique for designing privacy-preserving systems is "homomorphic encryption" (Gentry 2009). Homomorphic encryption allows a data processor to process encrypted data, i.e., users could send personal data in encrypted form and get some useful results, for example, recommendations of movies that online friends like, back in encrypted form. The original user can then again decrypt the result and use this without revealing any personal data to the data processor. This technique is currently still in its infancy; it does not scale yet to the large amounts of data stored in today's systems. However, if this could be made to work more efficiently, the results have the potential to be revolutionary (for privacy-preserving systems).

## Comparison and Critical Evaluation

As mentioned before, by far the easiest way to ensure that a system is privacy preserving is to not store or process any personal data. Of course, in practice, for many systems, this will not be possible. Such systems can use the techniques described in the previous section, but these each have their own problems and limitations. The section gives an overview of these issues.

One method for designing privacy-preserving systems is to use the various design principles and best practices such as ISO Standards, Privacy by Design, or the principles behind the EU Data Directive (transparency, purpose, proportionality, access, transfer). However, there are several problems with this. First of all, such rules and principles are typically rather vague and abstract. What does it mean to make a transparent design or to design for proportionality? The principles need to be interpreted and placed in a context when designing a specific – privacy-preserving – system. But different people will interpret the principles differently, and while this is useful in a legal setting where lawyers, prosecutors, and judges need enough freedom in their own interpretation of a particular situation (context!), this interpretation room is less helpful when one wants to design a system for a *specific* purpose: if several rules/guidelines are interpreted, the resulting system might not be privacy preserving because the interpretations might not fit together (are not composable). A more detailed design approach, with less room for interpretation, does not have this problem. Second, if one could agree on a specific, context-dependent, design of a privacy-preserving system, then that system still needs to be implemented. Implementation is another phase wherein choices and interpretations are made: system designs can be implemented in infinitely many ways. Moreover, it is very hard – for nontrivial systems – to verify whether an implementation meets its design/specification (Loeckx et al. 1985). This is even more difficult for nonfunctional requirements such as "being privacy preserving" or security properties in general (Warnier 2006).

Another privacy-preserving technique is anonymization of data. The idea is that by removing explicit links to individuals, the data can be safely processed for, for example, (medical) research purposes. The problem here is that it is very hard to anonymize data in such a way that all links with an individual are removed *and* the resulting anonymized data is still useful for research purposes. Researchers have shown that it is almost always possible to reconstruct links with individuals by using sophisticated statistical methods (Danezis et al. 2007) and by combining multiple databases (Anderson 2010) that contain personal information. Ultimately, how to address this issue is a trade-off between protecting privacy and advancing research. It suffices to say that even if databases with personal data are anonymized, access to them should remain restricted.

Dedicated software tools that provide anonymity of their users, such as Tor and Freenet, also have some problems. For example, Tor, the tool that allows anonymized communication and browsing over the Internet, is susceptible to an attack whereby, under certain circumstances, the anonymity of the user is no longer guaranteed (Back et al. 2001; Evans et al. 2009). Freenet (and other tools) have similar problems (Douceur 2002). Note that for such attacks to work, an attacker needs to have access to large resources that in practice are only realistic for intelligence agencies of countries.[3] However, there are other risks. Configuring such software tools correctly is difficult for the average user, and when the tools are not correctly configured, anonymity of the user is no longer guaranteed. And there is always the risk that the computer on which the privacy-preserving software runs is infected by a Trojan horse (or other digital pest) that monitors all communications (and knows the identity of the user). This is another example of the importance of context. Such tools can help to protect one's privacy (by providing anonymity), but that protection is never absolute.

In summary, numerous techniques exist for designing privacy-preserving systems, each with their own flaws. In practice, the most successful systems are designed for a specific purpose in a specific context. They typically combine several of the techniques described above.

## Experiences and Examples

Every system that stores or processes personal data has to be designed with privacy in mind. There are too many of such systems to discuss them here in any exhaustive manner. Instead, this section discusses in some detail one large system, the smart grid, as an example of what privacy issues arise in complex socio-technical systems and what mechanisms work and do not work in this context. Some examples of other systems that have similar issues are discussed at the end of the section.

[3]For example, the NSA can almost certainly indentify users of the TOR network. See https://www.eff.org/deeplinks/2012/03/https-and-tor-working-together-protect-your-privacy-and-security-online (retrieved 3/3/2012).

In the future power grid, the smart grid (Massoud and Wollenberg 2005), very large numbers of distributed (renewable) energy sources will be connected to the existing grid. These physically distributed generation installations (e.g., gas turbines, micro-turbines, fuel cells, solar panels, wind turbines) will be connected to the existing infrastructure. Integrated monitoring and control will make it possible to measure the effect on the grid, for example, to measure thermal stress caused by fluctuations in loading or fast transients due to DC to AC power conversion. Smart metering (McDaniel and McLaughlin 2009) devices, installed with consumers, enable applications such as peek prevention due to demand side management (Gellings and Chamberlin 1987) and the forming of virtual power stations (Ogston and Brazier 2009) by groups of consumers that sell their excess power (provided by solar or wind turbines) back into the grid. However, smart meters also store and process privacy-sensitive data, and they should be designed with care. Note the importance of context here: in a virtual power station, it is crucial that all consumption and production of electricity is carefully registered (using smart meters). However, this information is only stored and processed locally (within the virtual power station) and not shared with utility companies or other parties outside the virtual power station. Thus, smart metering itself does not harm one's privacy; only the specific context in which it is used might lead to a privacy violation.

Smart meter data can reveal many things about the members of a household, for example, it is easy to see from the power consumption pattern if the somebody is at home or how many people are a part of a household. More recently, researchers have shown that it is even possible to identify the movie that is being watched in a house, while other electrical appliances are in use, by solely observing the power consumption of the household (Greveler et al. 2012).

The privacy problems associated with smart metering have led to various outcomes. For example, legislators are – helped by special interest groups – becoming more aware of the problem, which has resulted in the blocking of legislation in 2009 by the Dutch Senate that was supposed to handle the mandatory role out of smart meters in the Netherlands (ESMA 2009). The main arguments against the plan were privacy concerns and a lack a choice for citizens if they wanted to participate (Fan et al. 2011). Electrical power companies have reacted to this by offering several different metering models for citizens, ranging from the old (off-line) system to smart meters that are under complete control of the power company (Boekema 2011). Consumers that give more control to the power companies receive a higher discount, in essence trading privacy for money.

That such a trade-off is not necessary is shown by privacy-preserving systems that try to serve both the interests of citizens (who, presumably, want privacy) and power companies (who want specific data on electricity use). A number of such privacy-preserving systems have been designed. Such systems are based on the techniques discussed in the previous section, such as anonymization (Efthymiou and Kalogridis 2010) or homomorphic encryption (Garcia and Jacobs 2011; Kursawe et al. 2011). Unfortunately, most of these systems are currently not operational. This is partly because of implementation issues but also because of incentives of power companies and end users. Power companies can make (more)

money by offering new services based on user's power consumption data or by selling (aggregated) data to governments and other organizations, and end users still do not ask for privacy and are willing to trade privacy for small monetary discounts. This shows again that, in essence, the specific context determines the success of privacy-preserving technologies: if someone can make money of privacy-sensitive data, it will usually happen (also see Facebook and Google). Legislation can help in such cases, but lack of enforcement remains a major issue.

Other examples of complex socio-technical systems that have similar privacy issues are electronic patient records in the health sector (Barrows and Clayton 1996; van 't Noordende 2010), public transport systems (Winters 2004; Garcia et al. 2008), electronic criminal records (Brazier et al. 2004; Warnier et al. 2008), and electronic social networks (Gross and Acquisti 2005; Rosenblum 2007). What all these systems have in common are as follows: (i) they store their information in digital form, (ii) they operate on the scale of countries or bigger, and (iii) different stakeholders have different incentives, roles, and interest in the system, in particular with regard to privacy. The first two points ensure that the systems can process more and more data automatically at ever-growing scales, which leads to ever more complex systems with more stakeholders (more organizations, countries, and people can become involved). This growing complexity is difficult enough to manage, but if the growing number of stakeholders, with different incentives (the context), is not taken into account, more and more of these systems will ultimately (inevitably!) fail to protect the privacy of its users.

## Open Issues and Future Work

One major (unsolved) issue in the design of privacy-preserving systems is that such systems are "dual use" (Atlas, and Dando 2006): they can be used to protect the privacy of citizens and dissidents, but they can also be used for illegal purposes such as terrorism and the distribution of child pornography. As the Freenet faq[4] states:

> **What about child porn, offensive content or terrorism?**
> While most people wish that child pornography and terrorism did not exist, humanity should not be deprived of their freedom to communicate just because of how a very small number of people might use that freedom.

This is a serious problem that has no realistic solution, but is too important to ignore (as the Freenet system does). Some privacy-preserving systems use key escrow schemes (Denning and Branstad 1996) for this: basically, the system allows the use of a master key that can "open" all encryption used in the system (and thus revealing the identity of criminal users). But it is unclear who should have access to the master key: the government? The United Nations? And if (when) it becomes

[4] http://freenetproject.org/faq.html#childporn (retrieved 3/3/2012).

known that such a key escrow scheme exists, nobody wants to use the system anymore, as, for example, the Clipper chip has shown (Froomkin 1995).

There are good guidelines and methodologies for the design of privacy-preserving systems, but there is still a lot of work to be done for the verification and validation of such systems: how do we know that a particular system indeed has the (privacy) properties we want? This remains an open research question.

## Conclusions

The multifaceted aspect of the concept privacy, with multiple stakeholders (with their own incentives), makes it difficult to design privacy-preserving systems. In general, "there is no golden bullet," a "one-size-fits-all" solution, to designing privacy-preserving systems. The particular context of the system needs to be taken into account. Even when new techniques, such as homomorphic encryption, become available, other (non-technical) issues such as context and incentives will at least be as important (if not more so).

## Cross-References


- ▶ Design for the Values of Accountability and Transparency
- ▶ Design for the Value of Trust


## References


Ackerman M, Darrell T, Weitzner D (2001) Privacy in context. Hum Comput Interact 16:167–176

Anderson RJ (2010) Security engineering: a guide to building dependable distributed systems. Wiley, New York

Annecharico D (2002) Notes & comments: V. Privacy after GLBA: online transactions: squaring the Gramm-Leach-Bliley act privacy provisions with the FTC fair information practice principles. NC Bank Inst 6:637–695

Atlas RM, Dando M (2006) The dual-use dilemma for the life sciences: perspectives, conundrums, and global solutions. Biosecur Bioterror Biodefense Strateg Pract Sci 4(3):276–286

Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. Comput Netw 54(15):2787–2805

Back A, Möller U, Stiglic A (2001) Traffic analysis attacks and trade-offs in anonymity providing systems. In: Sadeghi AR, Katzenbeisser S (eds) Information hiding. Springer, Berlin, pp 245–257

Barrows RC Jr, Clayton PD (1996) Privacy, confidentiality, and electronic medical records. J Am Med Inform Assoc 3(2):139–148

Berlin I (1958) Two concepts of liberty. Clarendon Press, Oxford

Birnhack M (2008) The EU data protection directive: an engine of a global regime. Comput Law Sec Rep 24(6):508–520

Boekema J (2011) Assessment of the implementation regulations for Smart Meters, TNO Technical Report, Delft, TNO-RPT-DTS-2011-00463-E

Brazier FMT, Oskamp A, Prins JEJ, Schellekens MHM, Wijngaards NJE (2004) Anonymity and software agents: an interdisciplinary challenge. AI Law 1–2(12):137–157
Cavoukian A (2009) Privacy by design. IPC, Ottawa
Cavoukian A (2010) Privacy by design: the definitive workshop. Identity Inf Soc 3(2):121–126
Clarke I, Sandberg O, Wiley B, Hong T (2001) Freenet: a distributed anonymous information storage and retrieval system. In: Federrath H (ed) Designing privacy enhancing technologies. Springer, Heidelberg, pp 46–66
Danezis G, Diaz C, Troncoso C (2007) Two-sided statistical disclosure attack. In: Proceedings of the 7th international conference on privacy enhancing technologies, Springer, pp 30–44
Denning DE, Branstad DK (1996) Key escrow encryption systems. Commun ACM 39(3):35
Dingledine R, Mathewson N, Syverson P (2004) Tor: the second-generation onion router. In: Proceedings of the 13th conference on USENIX security symposium, Washington DC, vol 13, p 21
Douceur J (2002) The Sybil attack. In: Peter D, Frans K, Antony R (eds.) Peer-to-peer systems. Springer, Berlin, pp 251–260
Efthymiou C, Kalogridis G (2010) Smart grid privacy via anonymization of smart metering data. In: First IEEE international conference on smart grid communications (SmartGridComm), New York, pp 238–243
EU Data Protection Directive (1995) Directive 95/46/EC of the European parliament and of the council of 24 Oct 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data
EU Proposal (2012) Proposal for a regulation of the European parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Brussels
ESMA (2009) Annual report on the progress in smart metering, Version 2.0
Evans NS, Dingledine R, Grothoff C (2009) A practical congestion attack on Tor using long paths. In: Proceedings of the 18th conference on USENIX security symposium, pp 33–50
Fan Z, Kulkarni P, Gormus S, Efthymiou C, Kalogridis G, Sooriyabandara M, Zhu Z, Lambotharan S, Chin W (2011) Smart grid communications: overview of research challenges, solutions, and standardization activities. IEEE Commun Surv Tutor 99:1–18
Feistel H (1973) Cryptography and computer privacy. Sci Am 228(5):15–23
Friedman B, Kahn Jr PH (2002). Human values, ethics, and design. In: Julie AJ, Andrew S (eds.) The human-computer interaction handbook. Lawrence Erlbaum Associates, Hillsdale, NJ, USA, pp 1177–1201
Friedman B, Kahn PH Jr, Borning A (2006) Value sensitive design and information systems. Hum Comput Interact Manag Inf Syst Found 5:348–372
Froomkin AM (1995) The metaphor is the key: cryptography, the clipper chip, and the constitution. Univ Pa Law Rev 143(3):709–897
Garcia FD, Jacobs BPF (2011) Privacy-friendly energy-metering via homomorphic encryption. In: 6th Workshop on Security and Trust Management (STM 2010) Lecture Notes in Computer Science, vol 6710. Springer, pp 226–238
Garcia FD, de Koning Gans G, Muijrers R, van Rossum P, Verdult R, Wichers Schreur R, Jacobs BPF (2008) Dismantling MIFARE classic. In: Jajodia S, Lopez J (eds) 13th European symposium on research in computer security (ESORICS 2008). Lecture Notes in Computer Science, vol 5283. Springer, pp 97–114
Gellings CW, Chamberlin JH (1987) Demand-side management: concepts and methods. The Fairmont, Lilburn
Gentry C (2009) Fully homomorphic encryption using ideal lattices. In: Proceedings of the 41st annual ACM symposium on theory of computing, ACM, pp169–178
Greveler U, Justus B, Loehr D (2012) Multimedia content identification through smart meter power usage proles. In: Gutwirth S, Leenes R, de Hert P, Poullet Y (eds.) Computers, privacy and data protection. Springer, Berlin
Gross R, Acquisti A (2005) Information revelation and privacy in online social networks. In: Proceedings of the 2005 ACM workshop on Privacy in the electronic society, ACM, pp 71–80

HIPAA (1996) Health insurance portability and accountability act of 1996
Hoare T (2003) The verifying compiler: a grand challenge for computing research. In: Proceedings of the 12th international conference on compiler construction. Springer, pp 262–272
Hone K, Eloff JHP (2002) Information security policy–what do international information security standards say? Comput Sec 21(5):402–409
Janger EJ, Schwartz PM (2001) The Gramm-Leach-Bliley act, information privacy, and the limits of default rules, Minnesota Law Review 86
Jones H, Soltren H (2005) Facebook: threats to privacy. Soc Sci Res 1:1–76
Juels A (2006) RFID security and privacy: a research survey. IEEE J Sel Area Commun 24(2):381–394
Kursawe K, Danezis G, Kohlweiss M (2011) Privacy-friendly aggregation for the smart-grid. In: Privacy enhancing technologies. Springer, pp 175–191
Liu C, Marchewka JT, Ku C (2004) American and Taiwanese perceptions concerning privacy, trust, and behavioral intentions in electronic commerce. J Glob Inf Manag 12:18–40
Loeckx J, Sieber K, Stansifer RD (1985) The foundations of program verification. Wiley, New York
Massoud SA, Wollenberg B (2005) Toward a smart grid: power delivery for the 21st century. Power Energy Mag IEEE 3(5):34–41
McDaniel P, McLaughlin S (2009) Security and privacy challenges in the smart grid. IEEE Sec Priv 7(3):75–77
Movius L, Krup N (2009) U.S. and EU privacy policy: comparison of regulatory approaches. Int J Commun 3:169–187
Nagenborg M (2009) Designing spheres of informational justice. Ethics Inf Technol 11:175–179
Nissenbaum H (2010) Privacy in context. Stanford University Press, Palo Alto
Ogston EFY, Brazier FMT (2009) Apportionment of control in virtual power stations. In: Proceedings of the international conference on infrastructure systems and services 2009: developing 21st century infrastructure networks, IEEE computer society, pp 1–6
PCI (2009) PCI security standards council, payment card industry (PCI) data security standard – requirements and security assessment procedures version 1.2
Phelps J, Nowak G, Ferrell E (2000) Privacy concerns and consumer willingness to provide personal information. J Public Policy Market 19:27–41
Rosen J (2010) The end of forgetting, The New York Times Magazine, July 25
Rosenblum D (2007) What anyone can know: the privacy risks of social networking sites. IEEE Sec Priv IEEE Comput Soc 5:40–49
Schorceman FD (ed) (1984) Philosophical dimensions of privacy: an anthology. Cambridge University Press, Cambridge
van den Hoven MJ (1999) Privacy or informational injustice? In: Pourcia L (ed) Ethics and information in the twenty-first century. Purdue University Press, West Lafayette, pp 140–150
van den Hoven J, Vermaas PE (2007) Nano-technology and privacy: on continuous surveillance outside the panopticon. J Med Philos 32(3):283–297
van 't Noordende G (2010) Security in the Dutch electronic patient record system, 2nd ACM Workshop on Security and Privacy in Medical and Home-Care Systems (SPIMACS), pp 21–31
Warnier ME (2006) Language based security for Java and JML. PhD thesis, Radboud University, Nijmegen
Warnier ME, Brazier FMT, Oskamp A (2008) Security of distributed digital criminal dossiers. J Softw 3(3):21–29, Academy Publisher
Warren SD, Brandeis LD (1890) The right to privacy. Harv Law Rev 4(5):193–220
Winters N (2004) Personal privacy and popular ubiquitous technology. In: Proceedings of Ubiconf, London
Zakaria N, Stanton JM, Sarkar-Barney STM (2003) Designing and implementing culturally-sensitive IT applications: the interaction of culture values and privacy issues in the middle east. Inf Technol People 16(1):49–75

# Design for the Value of Regulation

Karen Yeung

## Contents



**Keywords**

Regulatory instruments • Tools of government • Accountability • Responsibility • Regulation

## Introduction

Design has long been employed for regulatory purposes: by ancient civilizations (such as the design of Egyptian pyramids blocking burial shafts in order to discourage looters) through to contemporary communities (such as the use by digital media

K. Yeung (✉)
The Centre for Technology, Ethics, Law and Society (TELOS), The Dickson Poon School of Law, King's College London, London, UK
e-mail: karen.yeung@kcl.ac.uk



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_32

providers of "digital rights management" technology to prevent the unauthorized copying of digital data).[1] Identifying what counts as a "regulatory" purpose, however, is not entirely straightforward, largely due to the notorious lack of clarity concerning the meaning of the term "regulation." Suggested definitions range from narrow understandings of regulation as the promulgation of legal rules by the state and enforced by a public agency to extremely wide-ranging definitions which regard regulation as including all social mechanisms which control or influence behavior from whatever source, whether intentional or not.[2] Nonetheless, many scholars have increasingly adopted the definition of regulation proposed by leading regulatory theorist Julia Black as "a process involving the sustained and focused attempt to alter the behaviour of others according to defined standards or purposes with the intention of producing a broadly defined outcome or outcomes."[3] This definition captures the essential quality of regulation as systematic control and avoids a state-centric approach. Hence, it encompasses attempts by non-state institutions to shape social outcomes for defined purposes but is not so broad as to embrace the entire field of social science, thereby rendering regulation a relatively meaningless category.[4] At the same time, defining regulation in terms of intentional action aimed at affecting others provides the trigger for a plethora of concerns about its legitimacy, and it is this focus on intentionality which distinguishes regulatory scholarship from that of Science and Technology Studies (STS) scholarship which has long identified the ways in which artifacts can have social and political effects.[5]

Given the importance and ubiquity of regulation as a permanent feature of the governance of contemporary democratic economies,[6] it is hardly surprising that the field of "regulation" (or "regulatory governance" and its broader counterpart "governance") has become an established focus of scholarly analysis, drawing from a wide range of disciplinary orientations, including law, economics, political science, criminology, sociology, organizational theory, management studies, and other related social sciences.[7] Some scholars usefully portray and analyze regulation as a cybernetic process involving three core components that form the basis of

[1] Kerr and Bailey (2004), Ganley (2002)

[2] Black (2001), Daintith (1997)

[3] Black, *ibid* 142

[4] Black, *ibid*

[5] Winner (1980), Jelsma (2003), Akrich (1992)

[6] There are few spheres of economic activity that are not subject to some form of regulatory oversight and control, and daily news programs rarely pass without some mention of a significant regulatory decision, proposed regulatory reform, or allegations of some regulatory failure or scandal. Instances of alleged regulatory failure have been prominent throughout the late twentieth and early twenty-first century, including food safety (BSE in the 1990s and early 2000s), oil platforms (Piper Alpha in 1990, Deepwater Horizon in 2010), nuclear safety (Fukushima in 2011), and financial markets (Barings in 1995, the financial crisis post 2008)

[7] Baldwin et al. (2010)

any control system – i.e., ways of gathering information ("information-gathering"); ways of setting standards, goals, or targets ("standard-setting"); and ways of changing behavior to meet the standards or targets ("behavior modification").[8] Although design or technology can be employed at both the information-gathering (e.g., the use of CCTV cameras to monitor behavior) and behavior modification (e.g., offering candy to children to encourage them to act in desired ways) phases of the regulatory process, design-based regulation operates by preventing or inhibiting conduct or social outcomes deemed undesirable. It is the embedding of standards *into* design at the *standard-setting stage* in order to foster social outcomes deemed desirable (such as the incorporation of seat belts into motor vehicles to reduce the risk of injury to vehicle occupants arising from accidents and collisions) that distinguishes design-based regulation from the use of technology to facilitate regulatory purposes and processes more generally and which forms the focus of this paper.

Within regulatory studies literature, the use of design for regulatory purposes has not been the subject of extensive and comprehensive analysis, although particular kinds of design technologies have been the focus of considerable scholarly attention.[9] Nevertheless, two important themes can be identified within regulatory scholarship that may be of considerable assistance in interrogating contemporary debates concerning design for regulation: first, analysis of the tools or instruments that may be employed to implement regulatory policy goals and, secondly, debates concerning the legitimacy of regulation in particular contexts or the legitimacy of particular forms or facets of the regulatory enterprise. Both these themes will be explored in this paper through a discussion of the challenges associated with the effectiveness of design-based approaches to regulation and in the course of examining some of the controversies that have surrounded its use, with a particular focus on the implications of design for various dimensions of responsibility. In so doing, I will make three arguments. First, I will argue that design can be usefully understood as an instrument for implementing regulatory goals. Secondly, I will suggest that a regulatory perspective provides an illuminating lens for critically examining the intentional use of design to promote specific social outcomes by showing how such a perspective casts considerable light on their implications for political, moral, and professional accountability and responsibility. Thirdly, I will suggest that, because design can be employed for regulatory purposes (particularly in the case of harm mitigation technologies) without any need for external behavioral change on the part of human actors, Black's definition of regulation should be refined to bring all design-based instruments and techniques within the sphere of regulatory inquiry, rather than being confined only to those design-based approaches that intentionally seek to alter the behavior of others.

[8] Hood et al. (2001)

[9] Some of these applications are referred to below

## Understanding Design as a Regulatory Instrument

A well-established strand of regulatory literature is concerned with understanding the various techniques or instruments through which attempts might be made to promote social policy goals, primarily through the well-known policy instruments of command, competition, communication, and consensus – all of which seek to alter the external conditions that influence an individual's decision to act.[10] Consider the following strategies that a state might adopt in seeking to reduce obesity in the developed world, which is regarded by some as an urgent social problem.[11] It could enact laws prohibiting the manufacture and sale of any food or beverage that exceeds a specified fat or sugar level ("command")[12]; impose a tax on high-fat and high-sugar food products ("competition")[13]; undertake public education campaigns to encourage healthy eating and regular exercise[14] or attach obesity warning labels to high-fat and high-sugar foods ("communication")[15]; or offer specified privileges or benefits to high-risk individuals who agree to participate in controlled diet and exercise programs ("consent").[16] But in addition to all or any of the above strategies, a range of design-based (sometimes referred to as "code"-based or "architectural") approaches might be adopted, some of which are discussed below by reference to the subject in which the design is embedded (the "design subject").[17]

## Design Subjects

It can be helpful to classify design-based approaches to regulation by reference to design subject. These categories are not watertight, and many typically overlap so

[10]Morgan and Yeung (2007), Chap. 3

[11]Wadden et al. (2002)

[12]For example, New York City has banned the use of artificial trans fat in food service establishments in the city aimed at reducing the rate of heart disease (see Mello (2009))

[13]For example, Hungary has introduced a "fat tax" in an effort to combat obesity, and several US states have imposed an excise duty on sugar-sweetened beverages, partly motivated by a desire to combat obesity (see Cabrera Escobar et al. (2013))

[14]For example, in the UK, a national public awareness program including a public education campaign exhorting people to eat at least five portions of fruit and vegetables a day (the "5 A Day" program) was launched in 2002 to raise awareness of the health benefits of fruit and vegetable consumption and to improve access to fruit and vegetables (see http://webarchive.nationalarchives.gov.uk/+/www.dh.gov.uk/en/Publichealth/Healthimprovement/FiveADay/index.htm) (Accessed on 25 Nov 2013)

[15]For example, mandatory food labelling requirements imposed by EU law are considered by the European Commission as a central plank of the EU's obesity prevention strategy (see Garde (2007))

[16]For example, some US insurance companies and employers participate in wellness programs, pursuant to which employees are offered incentives in return for undertaking health-enhancing behaviors (see Mello and Rosenthal (2008))

[17]For a discussion of design-based approaches to regulation more generally, see Yeung (2008, 2016)

that often any given instrument might be placed in more than one category. Design instruments might also be readily combined.

### Designing Places and Spaces

When we think about design or architecture as a means for shaping behavior, we typically think of the ways in which places, spaces, and the external environment more generally may be designed to encourage certain behaviors while discouraging others. The crime prevention through environmental design (CPTED) approach to urban planning and design begins with the fundamental (and unsurprising) premise that our behavior is directly influenced by the environment we inhabit.[18] Hence, speed bumps can be installed in roads to prompt drivers to slow their speed, city centers can be pedestrianized to encourage greater physical activity, dedicated cycle lanes can be created to encourage people to cycle thereby encouraging better health and decreasing road congestion and air pollution from motor vehicles, and buildings can be designed with windows overlooking the street in order to increase the visibility of those passing along the street, discouraging crime and increasing the sense of security for street users and residents.

### Designing Products and Processes

Design may also be embedded in products or domestic and/or industrial processes in order to alter user behavior or their social impact. Hence, cone-shaped paper cups provided at water coolers discourage users from leaving their empty cups lying around because they cannot stand up unsupported, automatic cutoff mechanisms can be installed in lawnmowers to prevent the motor from running unless pressure is applied to the switch to prevent the lawnmower functioning unintentionally, and airbags can be fitted into motor vehicles to inflate on impact with another object in order to reduce the impact of the collision on vehicle occupants.[19]

### Designing Biological Organisms

The examples referred to above involve designing artifacts and environments that we encounter in our daily lives. But design-based approaches can also be extended to the manipulation of biological organisms, from simple bacteria through to highly sophisticated life-forms, including plants, animals, and, of course, human beings. So, for example, in seeking to reduce obesity, artificial sweeteners (such as aspartame or saccharin) instead of sugar might be used in processed foods, in order to reduce their calorific content; overweight individuals could be offered bariatric surgery in order to suppress their appetite and hence discourage food consumption

[18]Katyal (2002)

[19]Stier et al. (2007)

or anti-obesity medications (such as orlistat) might be provided to overweight individuals and others deemed to be at high risk of obesity.[20]

*Plants*: While crops bred for food production might be genetically modified to reduce the risks of obesity for developed world populations have not, to my knowledge, become a reality, the fortification of foods in order to enhance their nutritional value has a long pedigree. For example, niacin has been added to bread in the USA since the late 1930s, which is credited with substantially reducing the incidence of pellagra (a vitamin deficiency disease which manifests in symptoms including skin lesions, diarrhea, hair loss, edema, and emotional and psychosensory disturbance and, over a period of years, is ultimately fatal).[21] Plants can also be designed for a range of nonfood applications. For example, "pharming" involves genetically modifying plants and animals to produce substances which may be used as pharmaceuticals generating what advocates claim "an unprecedented opportunity to manufacture affordable modern medicines and make them available on a global scale."[22]

*Designing animals*: Genetic engineering for food production is on the cusp of extending beyond the bioengineering of plants to include more sophisticated life-forms including genetically modified fish (notably, salmon) designed for accelerated growth.[23] Several potential applications are also under investigation, including the introduction of genes to alter meat and milk composition to produce either leaner meat or enhanced antimicrobial properties of milk for newborn animals.[24] Biological engineering also offers considerable potential for reducing the prevalence and spread of infectious diseases. For example, an Oxford-based firm (Oxitec) has developed a genetically modified mosquito which it hopes will significantly reduce the spread of mosquito-borne disease such as dengue fever. Oxitec claims that these mosquitos have already been released for testing in Brazil, Malaysia, and the Cayman Islands, with test results indicating that mosquito numbers can be greatly reduced in a few months.[25] Similarly, genetically modified (transgenic) chickens that do not transmit avian influenza virus to other chickens with which they are in contact have been developed, thereby offering the prospect of employing this technique to stop bird flu outbreaks spreading within poultry flocks and thereby reducing risks of bird flu epidemics leading to new flu virus epidemics in the human population.[26]

[20]Of course, if a state proposed to implement any of these strategies, it would raise serious concerns about their legitimacy, particularly in relation to individuals who did not consent to the intervention, but these issues are beyond the scope of this paper (see Yeung (2015) *supra* n 7)

[21]Sempos et al. (2000)

[22]Paul et al. (2011)

[23]Krista et al. (2013)

[24]The European Food Safety Authority Panel has issued guidance on the environmental risk assessment of genetically modified animals, which includes insects, birds, fish, farm animals, and pets (EFSA Panel on Genetically Modified Organisms (2013))

[25]See BBC News (2013)

[26]The Roslin Institute, University of Edinburgh (2013)

*Designing humans*: Humans have a long history of seeking to interfere with their own biological processes and constitutions for a variety of social purposes. While the treatment of disease or its symptoms is clearly the primary motive for such interventions, there is no shortage of examples where technologies have been employed to alter human physiological function for various nonmedical purposes. Cosmetic surgery is perhaps the most well-known form of nontherapeutic surgical intervention, common in some developed economies through which individuals seek to "enhance" their physical appearance, including breast augmentation surgery, liposuction to remove fatty tissue, and skin tightening to reduce the appearance of wrinkles. Psychopharmacological approaches are also widely used to alter and lift mood and enhance mental cognition, particularly by students.[27] Human bioengineering has also made significant advances: preimplantation genetic testing and diagnosis could potentially be used as the basis for selecting embryos which display predispositions towards specific behavioral traits, and gene therapy might potentially be used to alter behavior through the repair or replacement of genes or the placement of a working gene alongside another faulty gene.[28] Advances in mechanical and digital engineering technologies and techniques have enabled the development of bionic technologies through which organs or other body parts can be replaced by mechanical versions, either mimicking the original function very closely or even surpassing it. For example, the cochlear implant is already widely used, and the rapid development of nanotechnology opens up the possibility for using extraordinarily powerful yet exceptionally small computer chips to enhance organ functions, including certain kinds of brain function.[29]

## Design Modalities

Design-based interventions can also be classified by reference to the mechanism through which they are intended to work. Consider again design-based mechanisms aimed at preventing and reducing obesity. First, the aim might be to *change individual behavior* by discouraging the purchasing and consumption of unhealthy foods and encouraging individuals to be more physically active. Thus product packaging could be designed to include ingredient lists and warning labels for foods with a high fat or salt content, and the installation of cycle lanes and pedestrianized city centers might seek to encourage greater physical activity. Product packaging can be understood as a form of "choice architecture," referring to the layout and social context in which individuals are provided with choices concerning their behavior that can be deliberately designed to encourage individuals to prefer some choices over others. One particular form of choice architecture that has attracted widespread publicity is the so-called "nudge" technique

[27]See, for example, Harris (2012) and Farah et al. (2004)
[28]Nuffield Council on Bioethics (2002)
[29]Foster (2006)

advocated by Thaler and Sunstein.[30] They define a nudge as "an aspect of choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives."[31] An oft-cited example is the image of a fly etched into urinals at Schiphol Airport that is designed to "improve the aim" because users subconsciously tend to aim at the fly etching, reducing the risk of spillage and hence helping to maintain the cleanliness of the facilities. The effectiveness of nudge techniques is claimed to rest on laboratory findings in experimental psychology which demonstrate that individuals systematically fail to make rational decisions, resorting instead to intellectual shortcuts and other decision-making heuristics which often lead to suboptimal decisions. The idea underpinning nudge strategies is that these "cognitive defects" can be harnessed through the shaping of choice architecture in order to encourage behaviors and social outcomes deemed desirable by the architect. Default rules and standards are considered to be particularly effective strategies for shaping behavior, seeking to harness the human tendency to "do nothing" and opt for the status quo.[32] For example, in the UK, Prime Minister David Cameron recently announced a policy initiative aimed at reducing the risks of children's access to pornography by securing the agreement of the six major Internet service providers to activate Internet filters against selected categories of content (not just pornography) unless the account holder (who must be over 18 years of age) *actively* opts to change the default setting to unfiltered Internet provision.[33]

Alternatively, design-based approaches may operate primarily by seeking to *prevent or reduce the probability of the occurrence of the undesired outcome*; hence, it might in future be possible to use preimplantation genetic diagnosis and selection to exclude embryos that have genetic markers for low metabolism thereby significantly reducing the risk of obesity faced by the resulting individual.[34] Finally, design-based approaches might seek to *mitigate the harm* generated by the relevant activity. Hence, the use of low calorie sweeteners as an alternative to sugar in manufactured food products enables consumers to continue consuming sweet-tasting carbonated drinks, without the high sugar content of sugar-sweetened variety. Some anti-obesity drugs work by reducing intestinal fat absorption by blocking fat breakdown and thereby preventing fat absorption, while others increase the body's metabolism, thus theoretically generating weight reduction without the need for the individual to change his or her dietary habits.[35]

[30]Thaler and Sunstein (2008)

[31]Ibid 6

[32]Ibid, 93

[33]The Rt Honourable David Cameron MP (2013)

[34]The Independent (2013)

[35]Nanoscience is currently being developed with a view to understanding how nanostructures contribute to the properties of food, thereby enabling food producers develop innovative ways of making similar products from different ingredients, for example, by removing most of the fat from ice cream without losing the smooth and creamy texture that consumers expect from that type of product (Ministerial Group on Nanotechnologies (2010))

Each of these "modalities of design" (as I have termed them) employs different mechanical logics in attempting to elicit the intended regulatory outcome. Although design-based approaches which seek to alter individual behavior, or which seek to prevent undesired social outcomes, fit comfortably within Black's definition of regulation, those which rely on harm mitigation do not because they need not generate any change to individuals' behavior. It might nevertheless be possible to interpret Black's definition in a way that would include such approaches. In particular, the relevant behavioral change which regulation seeks to elicit could extend beyond changes in the external behavior of individuals to include changes to the behavior and operation of an individual's *internal* physiological functioning (such as diet pills which increase metabolism) or the *behavior of material objects* in relation to each other (such as the impact of moving objects on shatterproof glass) or between living organisms and material objects (such as a cycle helmet's alteration of the impact of collision damage to the cyclist's head). But a more intellectually honest, and hence in my view preferable, approach would involve refining Black's original definition by removing the reference to behavioral change. Regulation would then be defined as "sustained and focused attempts intended process to produce a broadly defined outcome or outcomes directed at a sphere of social activity according to defined standards or purposes that *affect others*." This refined definition would manifest the three benefits identified by Black in support of her original definition by first, allowing inclusion of the purposive activities undertaken by non-state actors to shape social outcomes; secondly, avoiding a definition that is so extraordinarily broad that it essentially encompasses the whole of social science scholarship; and thirdly, it gives rise to the kinds of normative concerns about regulation and its legitimacy that have arisen in conjunction with the use of established techniques by those in authority to facilitate the achievement of regulatory goals by focusing on the intentional use of authority to *affect others*.[36] In this respect, it is worth emphasizing that the use of design-based techniques for self-regarding purposes by an individual (where, say, an overweight individual decides to embark on a course of diet pills) or by one individual for use by another (such as a doctor prescribing a course of diet pills to an individual patient in a clinical setting) does *not* amount to attempts to regulate because they aim to affect only one identifiable individual, rather than a *group* of individuals or organizations. However, if such measures were employed by, say, a public health agency in programmatic form (say by developing and implementing a nationwide program providing for the free supply and distribution of anti-obesity drugs to any person who met the criteria for eligibility), then this program *would* constitute a form of regulation. Defining regulation in terms of the targeting of groups or populations, rather than isolated individuals, is important because it is the exercise of authority over groups that lies at the foundation of concerns about regulatory legitimacy.[37]

[36]Black, above n 3

[37]Black (2008)

## Effectiveness, Rules, and Design

One of the most important issues raised within regulatory scholarship concerns the effectiveness of regulatory programs in achieving their intended outcome.[38] This section discusses the quest for regulatory effectiveness through design-based approaches by drawing on insights arising from studies of the use of rules in their traditional linguistic form. In particular, a core challenge faced by regulators lies in seeking to devise appropriate standards or rules that will provide clear and useful guidance to those they seek to regulate.[39] A rich and well-developed literature concerning the challenges associated with rules as guides for behavior demonstrates that traditional regulation in the form of a rule prohibiting specified activities backed by some kind of sanction for noncompliance (typically referred to as "command and control" regulation) can never be perfectly effective due to the inherent properties of rules.[40] First, rules are generalized abstractions that group together particular instances or attributes of an object or occurrence to build up a definition or category that forms the operative basis of the rule. Because these generalizations are inevitably simplifications of complex events, objects, or behaviors and because they are selective, sometimes properties will be included in the rule that are sometimes irrelevant, and some relevant properties will be left out.[41] Secondly, it is impossible to devise rules that cohere perfectly with their purpose; they will always be over-inclusive (catching situations that are irrelevant) or under-inclusive (failing to catch situations that ought to be included in order to secure the desired purpose). Even if there is a perfect causal match between the event and harm or regulatory goal, future events can develop in ways that the rule-maker has not, or could not have anticipated, so that the rule ceases to be perfectly matched to its goal. Thirdly, in seeking to provide guidance to those subject to the rule, clarity and certainty in their content and application will be of considerable importance. Yet the clarity of a rule is not solely a product of the linguistic text. It is also dependent upon shared understandings among those applying the rule (regulators, regulatees, institutions responsible for resolving disputes about the application of those rules). In other words, rules will invariably have what the English jurist and legal philosopher H.L.A Hart referred to as an "open texture," recognizing that although there will be clear cases that fall inside or outside the scope of a given rule, there will inevitably be a "penumbra of uncertainty" concerning its application to particular cases.[42]

At first blush, the use of design-based regulatory approaches may offer the promise of avoiding many of the inherent limitations of linguistic rules that are in large part a product of the indeterminacy of language. Yet a moment's reflection

[38] Yeung (2004)

[39] Baldwin et al. (2012), Chap. 14

[40] See, for example, Balwin (1995), Black (1997), Diver (1999), Schauer (1991)

[41] Black, *ibid*

[42] Hart (1961)

will soon reveal why design cannot deliver on this apparent promise. First, all design-based regulatory approaches rely on the embedding of standards into the fabric of the design target. Although the use of standards in *linguistic form* might be avoided through design-based approaches, the need for standards (and hence standard-setting) is not. Secondly, the problems of under- and over-inclusiveness remain whether or not rules take the form of linguistic constructs or features of material objects, biological organisms, or the built environment. So, for example, public outdoor seating is increasingly designed to make it impossible or uncomfortable for users to lie horizontally to discourage people from sleeping rough and thereby help promote equal access and enjoyment of public parks and amenities. Hence, a park's authority might install individual outdoor chairs or stools in public parks, rather than traditional bench-style seating. Although the former kind of seating will successfully discourage individuals from lying down, there may well be situations when the parks authority would have been willing to tolerate, or even encourage, an individual to lie on the bench in a particular instance, such as the injured jogger who sprains an ankle and seeks respite from her injury or a love-struck couple wishing to cozy up to each other to savor the park in full bloom, all of which might be regarded as activities that the parks authority, as regulator, would not wish to prevent.

Furthermore, standard-setting assumes even greater importance when design-based approaches to regulation are contemplated, particularly given the need to incorporate some kind of default standard into the design subject. The binary logic of technical design standards is not subject to the uncertainties arising from the inherent indeterminacy of language that plagues the use of linguistic rules. Nevertheless, some kind of default rule is needed to avoid operational failure or suspension in the event of unforeseen circumstances. For example, imagine that commercial passenger aircraft could be fitted with digital signal blocking devices to prevent passengers from using their mobile phones and other digital devices during aircraft takeoff and landing, in order to ensure that the aircraft's communication systems are not interfered with during these crucial flight periods. Provision would need to be made for any unrecognized signal to be dealt with as either "permissible" (thereby allowing the signal to continue transmission) or as a "violation" (thereby automatically blocking transmission). Such a default standard avoids the need for human interpretation, thereby ensuring that a regulatory response will obtain for every situation. Yet it cannot ensure that the response will be aligned with the regulator's underlying policy objectives in each and every case. If the default device is programmed to block any unrecognized signals, this might generate a minor inconvenience to those who find that their portable entertainment systems will not operate, but the consequences would be considerably more serious for a passenger suffering from Parkinson's disease who relies upon deep brain stimulators for the treatment of his neuropathic pain and tremor control.[43]

[43]Lyons (2011)

Unlike linguistic rules, design-based instruments can be self-executing so that once the standard embedded within the design object has been reached, the response is automatically administered, thereby forcing a particular action or set of actions to occur (hence, I refer to them as "action-forcing" designs).[44] For example, digital locks automatically prevent the unauthorized copying of "locked" digital data – there is no need for an administrator or other official to administer the power of exclusion once the digital lock is in place. By contrast, the violation of linguistic rules (such as a "no parking" sign) cannot be sanctioned unless and until compliance with the rule is actively monitored and enforced. Not only does this require human personnel to undertake monitoring and enforcement action against suspected noncompliance but it also requires – at least in democratic societies – a set of enforcement institutions to oversee and administer the lawful and proper application of sanctions. Linguistic rules require interpretation, enforcement, and sanction through human interaction, in which a discrete set of factual circumstances must be interpreted and applied by human agents and an appropriate response identified and administered.

Because rule enforcement is a resource-intensive activity, many legal violations that might otherwise have been proved to the standards required by a court of law might nevertheless go unpunished, particularly those of a fairly minor nature including trivial road traffic violations. In this respect, design-based instruments that avoid the need for human enforcement (which are increasingly referred to as "autonomous technologies") appear to offer a considerable advantage over their more traditional rule-based counterparts, obviating the need for human and institutional enforcement resources while offering consistent and immediate application. Yet socio-legal scholars have amply demonstrated that the sensitive and judicious exercise of discretion by enforcement officials serves a vital role, enabling regulatory rules to be applied in a manner that conforms with their underlying "spirit" or policy objective, rather than insisting on strict compliance where this is judged to be counterproductive.[45] Hence, a parking inspector may exercise her discretion not to issue an infringement notice against a vehicle parked temporarily in a "no parking zone" to allow the driver to unload heavy items of furniture for the purpose of transferring them into the adjacent house when this does not seriously inhibit the free flow of traffic. In other words, within traditional rule-based regulatory regimes, inescapable problems of inclusiveness and determinacy that arise at the rule-setting stage can be addressed at the enforcement stage through sensitive interpretation and application. Although human involvement in the application of rules can be a source of inconsistency and error, it also provides the vehicle through which the limitations of rules can be overcome in concrete contexts.

[44]For a discussion and critique of self-enforcement in the context of "tethered" digital appliances, see Zittrain (2007)

[45]See, for example, Hawkins (1984, 2002), Hutter (1997), Grabosky and Braithwaite (1985)

## Design-Based Regulation, Agency, and Responsibility

Another central theme arising in debates concerning the legitimacy of regulatory regimes focuses on the accountability of regulatory agencies and institutions. These concerns are rooted in the need for mechanisms through which those in positions of authority whose decisions and activities have the power to affect others should, at least within liberal democratic politics, be held appropriately accountable for their actions. Within legal and political studies literature, the term "accountability" is often used as a synonym for many loosely defined political desiderata, such as good governance, transparency, equity, democracy, efficiency, responsiveness, responsibility, and integrity.[46] Mark Bovens suggests that, broadly speaking, scholarly analysis of accountability adopt two rather different conceptions: either as a virtue or virtuous behavior or as a mechanism or a specific social relation that involves an obligation to explain and justify conduct from the actor (the accounter) to a forum, the account holder, or accountee.[47] It is this second sense of accountability that is typically the focus of discussion in debates about regulatory legitimacy. On this understanding, accountability usually involves not just the provision of information about performance but also the possibility of debate, of questions by the forum and answers by the actor, and eventually of judgment of the actor by the forum. Furthermore, judgment also implies the imposition of formal or informal sanctions on the actor in case of poor or unacceptable performance or rewards in cases of adequate or superior performance.[48] So conceived, obligations of accountability can be understood as flowing from the position of responsibility which the accounter occupies in relation to the account holder. These obligations extend beyond offering an explanation of one's actions but also require the accounter to take responsibility for the impact and effect of their decisions on others, including an obligation of responsiveness – to respond to the needs and demands of those to whom they are required to account, to make amends when things go wrong, or to make adjustments or changes to their proposed course of action when account holders so demand. Accountability can therefore be conceived as a product of the *position of responsibility* occupied by the designated account holder.[49] Used in this sense, responsibility has a temporal element that looks in two directions.[50] Notions of accountability and answerability look backwards to conduct and events of the past: what Peter Cane refers to as "historic responsibility." In contrast, "prospective responsibilities" are future oriented, concerned with establishing obligations and duties – and are typically directed towards producing good outcomes ("productive responsibilities") and preventing bad outcomes ("preventative responsibilities").[51]

[46]Bovens (2010) and the literature cited therein
[47]Ibid, 949–951
[48]Ibid
[49]Gardner (2006)
[50]Cane (2002)
[51]Ibid

Regulators, like engineers, are typically understood as responsible in both senses, although the focus of much regulatory literature has been on the historic rather than prospective responsibilities of regulatory officials. Moreover, notions of responsibility and accountability can be understood from a variety of perspectives, depending upon the particular dimensions or kind of decision-making judgment required and considered salient: be it political, professional, financial, moral, scientific, administrative, or legal, to name but a few. In the following section, I shall consider some of the implications of design-based approaches to regulation on three different dimensions of responsibility: political, professional, and moral. Although a varied range of concerns have been expressed in relation to each of these dimensions of responsibility, they are ultimately rooted in the account holders' discretionary power to trade off competing considerations and the need to ensure that – at least in regulatory contexts – those to whom decision-making authority is entrusted should be held accountable and responsible for the consequences of their judgments.

## Design-Based Regulation and Political Responsibility

The regulation of social and economic activity in the last three to four decades in many industrialized economies has been accompanied by the increasing popularity of the independent regulatory agency as an institutional form.[52] Such agencies are typically established by statute and endowed with statutory powers, yet are typically expected to operate at arm's length from the government rather than being subject to regular ministerial direction. Although this institutional form has a long history, it was the proliferation of utility regulators following the privatization of state-owned natural monopolies and their subsequent regulation by independent regulatory agencies that began to attract scholarly attention.[53] A number of benefits are claimed to be associated with the use of independent agencies in carrying out regulatory functions: their capacity to combine professionalism; operational autonomy; political insulation; flexibility to adapt to changing circumstances, continuity, and hence capacity to adopt a long-term perspective rather than being subject to the vagaries of the electoral cycle; as well as policy expertise in highly complex spheres of activity.[54] Yet they have also attracted considerable criticism, primarily on the basis that such agencies lack democratic legitimacy and are not adequately accountable for their decisions.[55] Because their decisions have a differential impact on individual and group interests, and frequently require them to make trade-offs between competing values and principles, the decisions of regulatory agencies can be understood as having political dimensions, underlining the need for

[52]Levi-Faur (2005)

[53]Black (2007)

[54]Levi-Faur, above n 51; Levy and Spiller (1996)

[55]See, for example, Graham (1998), Baldwin (1996), Yeung (2011a), Scott (2000)

mechanisms to promote democratic accountability in regulatory decision-making. Although regulatory agencies are typically subject to specific mechanisms of accountability, such as public reporting obligations to the parliament and accountability to the courts by way of judicial review of agency decision-making, those appointed to lead and manage such agencies are not elected, nor are they directly accountable to national legislatures or subject to direct ministerial control; hence, it is not surprising that complaints are frequently made that regulatory agency decisions lack democratic legitimacy.[56]

If political accountability is considerably weakened by the transfer of decision-making authority from democratically elected ministers to independent regulatory agencies when they employ traditional command-based approaches to regulation, there are reasons to believe that the use of design-based instruments exacerbates these weaknesses. Such concerns have been particularly potent in debates concerning the use of code-based approaches to regulating the Internet. Cyberscholar and constitutional lawyer Lawrence Lessig has famously claimed that, within cyberspace, "code is law," observing how software code operates to restrict, channel, and otherwise control the behavior of Internet users.[57] Two related sets of concerns have arisen focused upon the potential for code-based regulation to undermine democratic accountability. First, it is claimed that when employed by the state, code-based regulation may antagonize several constitutional principles: its operation may be opaque and difficult (if not impossible) to detect, thereby seriously undermining the transparency of regulatory policy; the lack of transparency diminishes the accountability of those responsible for installing and operating code-based controls; and as is the extent to which affected individuals may participate in the setting of such controls before they are installed or to challenge or appeal against such policies after they have been imposed.[58] As a result, both authoritarian and libertarian governments alike can enforce their wills much more easily than they could through more traditional command-based approaches, yet without the knowledge, consent, or cooperation of those they govern.[59] Secondly, when code-based approaches are employed by non-state actors in pursuit of private goals, particularly by extraordinary powerful commercial entities such as Google, Amazon, and Facebook, this may subvert or override the legislatively authorized balance of values, profoundly altering the balance of power between governments and the governed.[60]

Similar kinds of concerns have been targeted at the use of "nudge" techniques to encourage individuals to behave in ways deemed desirable by the "nudger." Although controversy over the legitimacy of such techniques has been wide-ranging, for present purposes it concerns about the transparency of such techniques

[56]Ibid

[57]Lessig (1999) drawing on the insight provided by Joel Reidenberg (Reidenberg (1998))

[58]Citron (2008)

[59]Lessig, *ibid*

[60]Ibid

that are of particular salience.[61] For example, open Internet campaigners have criticized (and ridiculed) the UK government's default Internet filtering policy aimed at reducing children's access to Internet porn, based on concerns about the consequent loss of transparency, accountability, and due process entailed by the policy, questioning whether the ISPs implementing the Internet filters would be responsible for incorrect blocks and financially liable to those suffering economic loss as a result.[62] Even Thaler and Sunstein acknowledge that the nudge techniques they advocate can be likened to subliminal advertising in that some nudges may be insidious, empowering governments to "manoeuvre people in its preferred directions, and at the same time provide officials with excellent tools by which to accomplish this task."[63] They therefore propose a form of John Rawls's publicity principle as a limit on the legitimate use of nudges prohibiting governments from adopting policies that they would not be able or willing to defend publicly on its own grounds.[64] But judging from the response of the US and UK governments to recent revelations following whistle-blower and former US intelligence contractor Edward Snowden's disclosure of the extent to which US and UK intelligence agencies have been monitoring digital and other forms of communications of their own citizens without their consent, openly defending their actions with little or no apparent embarrassment, this proposed principle is unlikely to provide much of a safeguard.[65]

## Design-Based Regulation and Professional Responsibility

Concerns about the ways in which design-based approaches to regulation may entail the exercise of political judgment involving trade-off between conflicting values and interests, yet are not subject to effective mechanisms to ensure that those exercising such judgments are rendered responsible and accountable for doing so, have direct parallels in professional contexts. In particular, debates about the appropriate use of design to promote the goal of avoiding unintended errors by medical practitioners in the provision of healthcare highlight how the use of design

[61]See, for example, Bovens (2008), White (2010), Yeung (2012), Rizzo and Whitman (2009), Schlag (2010)

[62]The UK Open Rights Group, for example, argues that because this measure has been introduced without any legislative authority, the public appears to have no recourse when things go wrong, and there will be no one to pressurize (see the Open Rights Group Campaign (2013))

[63]Thaler and Sunstein above n 30, 244

[64]Ibid 244–245

[65]The disclosures by Edward Snowden and the responses of various heads of the government have received very extensive media coverage. On the response of the US administration, see, for example, K Connolly "Barack Obama: NSA is not rifling through ordinary people's emails," *The Guardian*, London, 19 June 2013; on the response of the UK administration, see, for example, S Jenkins "Britain's response to the surveillance scandal should ring every alarm bell," *The Guardian*, 4 November 2013

to foster laudable goals such as ensuring patient safety in the practice of medicine and public health provision may have significant and potentially troubling implications for professional agency and responsibility.

Traditionally, professional and legal standards have been relied on to secure patient safety, based on an agent-centered approach to regulation. In particular, the regulation of the medical profession has historically taken the form of professional "self-regulation" in which an association of medical professionals established in corporate form seeks to exert control over individual doctors by controlling entry to the profession through a system of licensing for approved agents, largely leaving individual members to exercise their own agency in behaving well and complying with the association's code of conduct.[66] Such an approach relies heavily on individuals internalizing and cooperating with the collective norms of the professional group. The effectiveness of this mechanism has been the focus of powerful critiques. The most trenchant critiques express skepticism of the two major claims that underpin faith in professional agency: the expertise claim (doctors have specialist, distinctive knowledge and skills that are inscrutable to others) and the moral claim (doctors will reliably act in the interests of their patients and apply their expertise diligently to secure those interests).[67] Thus, patient safety failures are seen as inevitable because the character, conscientiousness, competence, and good motives of individual agents cannot be satisfactorily relied on to ensure patient welfare, and the profession as a corporate body is incapable of ensuring that its members comply with the appropriate standards. Another, more sympathetic critique also regards reliance on individual agents as an ineffective means for ensuring patient safety but derives from rather different assumptions. It is based on recognition that humans inevitably make errors, so that not only is human agency ineffective as a means for avoiding or reducing error but it is also unfair and unhelpful to doctors.[68] It therefore advocates a focus on the conditions under which individuals work, emphasizing the systemic and environmental causes of error, with the aim of constructing defenses to avert or mitigate them.[69] This gives rise to a "systems-based" approach, which seeks to prevent errors through careful system design, rather than a reliance on the competence and conscientiousness of individual agents.[70] Hence, a systems-based approach to patient safety emphasizes the ways in which the architecture or design of healthcare settings can be "mistake-proofed" thereby making it impossible or considerably more difficult for practitioners to cause harm.

One form of design-based mistake-proofing involves the use of action-forcing design. Wrong-route drug administration is a commonly used example of the kind of behavior that could be avoided by action-forcing design. This is often seen as an

[66]Rostain (2010)

[67]Friedson (1973)

[68]Merry and McCall Smith (2001)

[69]Reason (2000)

[70]Department of Health (2000), Kohn et al. (2000)

egregious patient safety error. Importantly, it typically occurs unintentionally (either when an error in planning is made, i.e., someone plans to give a patient a drug through the wrong route, without realizing the wrong route, or when an error of execution is made – i.e., someone does not plan to give a patient through the wrong route but accidentally does so). Both the action and outcome are unintended – the results of lapses. A classic example is administration of vincristine (chemotherapy) via the intrathecal route (the spine) which has very serious (usually fatal) consequences.[71] Redesigning equipment could mean that it would no longer be possible to connect a normal hypodermic syringe to a spinal device, making it impossible inadvertently to administer the drug intrathecally. Such a solution is likely to be welcomed across various stakeholders and interests in the same way as the redesign of anesthetic equipment to prevent nitrous oxide being administered instead of oxygen has enjoyed widespread legitimacy.

But, the simplicity, appeal, and likely effectiveness of an action-forcing design solution conceal underlying ethical controversy in the healthcare context, where definitions of risk, morality, and error are often highly contested and where professional agency has traditionally had an important role. Wrong-route drug administration is universally regarded as a serious medical error. Yet in other circumstances where there is contestation in defining what actions constitute an error, who should own the definition and the conditions in which such actions should be prevented, design-based regulation becomes considerably more problematic. Consider, for example, the use of design-based approaches to prevent the reuse of medical devices in order to reduce infection risk arising from ineffective sterilization or damage to reusable equipment. In many countries, official policy is that any devices designated as single-use ("SUD") must not be reused.[72] On one view, any reuse of a SUD would be an error. Yet the reuse of medical devices labeled "single-use only" by manufacturers is highly contentious; there is little consensus on how far reuse constitutes a genuine safety risk for many devices. Some argue that when appropriate precautions are taken, reuse can very often be justifiable for many such devices where there is a very low risk associated with reuse and also that there are good environmental and economic reasons to do so.[73] Practitioners may perceive that manufacturers are overcautious and self-serving in their instructions cautioning against reuse. Hence, they may intentionally reuse equipment even though they intend no harm, especially if this prevents waste and allows more patients to benefit from limited resources.[74]

Yet manufacturers have increasingly been designing single-use devices in such a way as to render them auto-disabling, thereby preventing reuse (e.g., single-use needles, self-blunting needles). Seen in light of considerable contestation about the

[71]For example, British teenager Wayne Jowett died in Nottingham, England, in 2001 following intrathecal administration of vincristine (Toft (2001))

[72]For example, Medicines and Healthcare Devices Regulatory Authority (2006)

[73]Kwayke et al. (2010)

[74]Smith et al. (2006), Dickson

legitimacy of reusing medical devices designated as single-use only, these manufacturing innovations take on a much more problematic guise. Rather than being neutral, value-free interventions, action-forcing design imposes an action that favors one particular judgment about what constitutes an "error" and what constitutes "safe" medical practice. Unlike wrong-route drug administration, the lack of consensus about the reuse of single-use devices means that action-forcing design that prevents reuse may encode a technical notion of risk that may appear objective but serves to obscure normative and programmatic commitments on the part of designers. Not only does this crowd out doctors' professional discretion and accountability for making value judgments about the appropriate balance between patient safety, economic prudence, and environmental sustainability, but it may also serve to exclude stakeholder participation in the setting of standards and allow penetration of commercial and other interests for which there is little transparency or public accountability.[75]

## Design-Based Regulation and Moral Responsibility

The need to exercise individual judgment in trading off competing values and concerns also has important moral analogues. Just as the turn to design-based approaches to regulation has significant implications for political and professional judgment and responsibility, it also raises potentially more profound implications for our understanding and practice of moral judgment and responsibility. Concerns about the potential for design-based approaches to erode or otherwise undermine moral responsibility are evident in a range of different literatures from various disciplines when used to shape or channel social behavior. For example, leading criminologist David Garland refers to "situational crime prevention" as a "set of recipes for steering and channelling behaviour in ways that reduce the occurrence of criminal events. Its project is to use situational stimuli to guide conduct towards lawful outcomes, preferably in ways that are unobtrusive and invisible to those whose conduct is affected."[76] While Garland explains the political appeal of these strategies for governments by offering a more immediate form of security to potential victims, and one that can be increasingly commercialized through the involvement of private sector providers, Duff and Marshall worry that such techniques may express a lack of respect for individuals, implying that individuals are incapable of responding to appeals to moral reasoning or exercising self-control and restraint.[77] In a different but related vein, applied ethicists have raised concerns about allowing the use of technological approaches to enhancing individual traits and capabilities when used collectively to promote nonmedical goals. So, for example, Allan Buchanan argues that the quest for economic growth is likely to

[75]For a fuller analysis, see Yeung and Dixon-Woods (2010)

[76]Garland (2000)

[77]Duff and Marshall (2000)

result in state support for the use of human enhancement technologies that improve industrial productivity while advances in neuroscientific knowledge have provoked a resurgence of interest in "biological approaches to crime control."[78] Although these controversies are of relatively recent origin, they can also be understood as offering contemporary applications of much more long-standing controversies about the moral and ethical legitimacy of using design ostensibly to shape human progress and flourishing such as concerns surrounding the legitimacy of water fluoridation to reduce population level tooth decay,[79] state-sponsored vaccination programs to prevent and limit the spread of infectious disease,[80] and state-sponsored eugenic programs aimed at breeding a superior species.[81]

Taken together with concerns expressed about the potential for design-based approaches to undermine democratic responsibility,[82] these apparently disparate critiques reflect a common set of anxieties that the use of design-based approaches for influencing human affairs could threaten the moral and social foundations to which individual freedom, autonomy, and responsibility are anchored. These foundational concerns have been alluded to by legal scholars Roger Brownsword and Ian Kerr. Both fear that, when used on a cumulative and systematic basis, such approaches may fatally jeopardize the social foundations upon which moral community rests. Hence, Brownsword fears that the use of action-forcing design (which he terms "techno-regulation") entails more than a loss of moral responsibility but in a direct and unmediated way excludes moral responsibility because individuals who are forced by the designed environment that they inhabit to act in particular ways can no longer be regarded as morally responsible.[83] Because action-forcing design deprives agents of the opportunity to choose how to behave, it deprives them of the opportunity to exercise moral judgment. Similarly, Ian Kerr foreshadows the potential social consequences of a more generalized strategy that relies upon action-forcing technologies (which he refers to as "digital locks"). For him, such approaches may stultify our moral development by eliminating the possibility for moral deliberation about certain kinds of action yet leave "no room for forgiveness."[84] He therefore fears that "a successful, state-sanctioned, generalized deployment of digital locks actually impedes the development of moral character by impairing people's ability to develop virtuous dispositions, thereby diminishing our well-being and ultimately undermining human flourishing."[85]

Although Brownsword's concerns that action-forcing technologies eliminate moral agency are, in my view, overstated – at least in circumstances where agents

[78]Raine (2013)
[79]Connett et al. (2010), Peckham (2012)
[80]Colgrove (2006)
[81]Romero-Bosch (2007)
[82]See section "Design-based Regulation and Political Responsibility" above
[83]Brownsword (2006)
[84]Kerr (2010)
[85]Ibid

have an adequate range of alternatives to act in ways that they consider morally right or wrong – nevertheless both Brownsword and Kerr are rightly fearful of the implications for our moral foundations of a systematic shift in favor of such measures for implementing public policies. Such a shift is considerably more likely to arise incrementally and cumulatively, rather than through a single highly visible change in regulatory approach at a discrete point in time, and hence much more likely to escape public notice. Not only do we need to reflect carefully on the moral risks posed by particular design-based regulatory technologies considered in isolation, but particular vigilance is needed in attending to the *systemic* moral risks associated with design-based regulation, including the articulation of an analytical framework that can assist in conceptualizing, analyzing, and debating the collective shift towards design-based approaches to regulation.[86]

## Conclusion

This paper has shown how design can be employed as an instrument of regulatory control, used intentionally by state and non-state actors in particular contexts for the purposes of producing broadly defined outcomes which affect others. Because design can be employed for regulatory purposes without necessarily seeking to elicit a change in the external behavior of others, particularly in the case of harm mitigation technologies, I have suggested that Julia Black's definition of regulation should be refined in a way that will allow such design-based approaches to be included within the regulatory scholar's field of vision. Drawing upon two significant themes and literatures within regulatory scholarship, the first concerning regulatory tools and instruments and the second concerned with the accountability and legitimacy of regulatory agencies, I have demonstrated how a regulatory perspective can illuminate important ethical debates that may arise when design is employed for regulatory purposes. For regulatory authorities, the attractions of design lie in their self-enforcing capacities, thereby avoiding both the expense and potential for the improper exercise of authority by individuals entrusted with the task of enforcing regulatory rules while securing the swift and effective achievement of regulatory goals. Where there is strong consensus about the kinds of behaviors and activities considered undesirable, then appropriately formulated design-based interventions may deliver considerable benefits and command widespread acceptance by regulators and those they regulate.

But even where such consensus exists, I have shown how difficulties associated with the setting of standards in traditional linguistic, rule-based form are likely to be exacerbated, rather than diminished, through the incorporation of regulatory standards into the fabric of design, at least in circumstances where unforeseen

[86]For one suggested approach, drawing on common pool resource theory and the "tragedy of the commons" (see Yeung (2011b))

circumstances arise which have not been contemplated by designers. Nor are attempts to shape social outcomes through design, rather than more traditional policy approaches, likely to overcome or avoid controversies associated with the accountability and responsibility of regulators. Rather, because design-based approaches to regulation seek to encode standards into the fabric of design, some mechanism may be needed to resolve the inevitable trade-offs between conflicting values and interests in concrete contexts. Design makes it possible to encourage or compel actions or outcomes deemed by regulators as desirable in ways that both obscure and deepen concerns about their political accountability. When used to guide and shape professional judgment, including clinical decision-making by doctors, the use of design-based approaches to promote "good" medical practice can be controversial, at least in situations where there is a lack of consensus about what constitutes "good" clinical practice yet the design operates to preclude certain kinds of activities (such as the reuse of single-use medical devices) and may both complicate and erode the professional accountability of clinicians. Finally, and perhaps most worryingly, the use of design-based approaches to regulation has the potential to undermine moral responsibility and accountability, at least in circumstances where the turn to design-based approaches to regulation becomes so systematic and routine that it results in a significant erosion of the extent to which individual agents are left free to make their own moral judgments and act upon them accordingly.

A significant theme emerging in recent philosophy of technology literature focuses on the interface between responsibility and engineering, highlighting how engineering and technology increasingly shape the context of human actions, and therefore informs and influences how responsibility in both its prospective and retrospective senses is understood and distributed.[87] In common with regulatory accountability scholarship, this literature reflects a shared concern that those who wield the power to trade off competing values that may affect the rights, interests, and legitimate expectations of others should be appropriately held to account to those affected others, enabling them to seek redress, appeal, or prompt reconsideration of past decisions or prospective policies in light of their feedback and experience. Just as scholars of engineering ethics have drawn attention to the trade-offs between values that may be involved in the engineering design process, so also have regulatory scholars sought to identify and evaluate the extent and adequacy with which regulators of all stripes, within a varied range of institutional and policy contexts, are held accountable and responsible for the way in which they have traded off conflicting values and interests in carrying out their regulatory duties. These debates are likely to intensify rather than subside, as our technological knowledge and capacity continues to advance, thereby opening up the possibility of more powerful, precise, and invasive regulatory design strategies than our forefathers could possibly have imagined.

[87]Doorn and van de Poel (2012)

## References


Akrich M (1992) The description of technical objects. In: Bijker W, Law J (eds) Shaping technology. MIT Press, Cambridge, MA

Baldwin R (1995) Rules and government, Oxford socio-legal studies. Clarendon, Oxford

Baldwin R, Cave M, Lodge M (2010) Introduction: regulation – the field and the developing agenda. In: Baldwin R, Cave M, Lodge M (eds) The Oxford handbook of regulation. Oxford University Press, Oxford

Baldwin R, Cave M, Lodge M (2012) Understanding regulation, 2nd edn. Oxford University Press, New York

Balwin R (1995) Rules and government. Oxford University Press, New York

Black J (1997) Rules and regulators. Clarendon, Oxford

Black J (2001) Decentring regulation: understanding the role of regulation and self-regulation in a 'post-regulatory' world. Curr Leg Probl 54:103

Black J (2007) Tensions in the regulatory state. Public Law 58

Black J (2008) Constructing and contesting legitimacy and accountability in polycentric regulatory regimes. Regul & Govern 2(2):137–164

Bovens L (2008) The ethics of *Nudge*. In: Grune-Yanoff T, Hansson SO (eds) Preference change: approaches from philosophy, economics and psychology. Springer, Dordrecht, Heidelberg

Bovens M (2010) Two concepts of accountability: accountability as a virtue and as a mechanism. West Eur Polit 33:946

Brownsword R (2006) Code, control, and choice: why east is east and west is west. Legal Stud 25:1

Cabrera Escobar MA et al (2013) Evidence that a tax on sugar sweetened beverages reduces the obesity rate: a meta-analysis. BMC Public Health 13:1072

Cane P (2002) Responsibility in law and morality. Hart Publishing, Oxford and Portland, Oregon, UK, 31

Citron DK (2008) Technological due process. Wash Univ Law Rev 85:1249

Colgrove J (2006) State of immunity: the politics of vaccination in twentieth century America. University of California Press, Berkeley

Connett P, Beck J, Spedding Micklem H (2010) The case against fluoride: how hazardous waste ended up in our drinking water and the bad science and powerful politics that kept it there. Chelsea Green Publishing, White River Junction, VT

Daintith T (1997) Regulation. In: Buxbaum R, Madl F (eds) International encyclopedia of comparative law, vol XVII, State and economy. JCB Mohr (Paul Siebeck), Tübingen

Department of Health (2000) An organisation with a memory. Department of Health, London

Dickson DE (1999) Rapid response to: controversy erupts over reuse of 'single use' medical devices. http://bmj.com/cgi/eletters/319/7221/1320

Diver CS (1999) The optimal precision of administrative rules. In: Baldwin R, Hood C, Scott C (eds) A reader on regulation. Oxford University Press, New York

Doorn N, van de Poel I (2012) Editor's overview: moral responsibility in technology and engineering. Sci Eng Ethics 18:1

Duff R, Marshall S (2000) Benefits, burdens and responsibilities: some ethical dimensions of situational crime prevention. In: von Hirsch A, Garland D, Wakefield A (eds) Ethical and social perspectives on situational crime prevention. Hart Publishing, Oxford, p 17

EFSA Panel on Genetically Modified Organisms (2013) Guidance on the environmental risk assessment of genetically modified animals. EFSA J 11:3200

Farah MJ et al (2004) Neurocognitive enhancement: what can we do and what should we do? Nat Rev Neurosci 5:421

Foster KR (2006) Engineering the brain. In: Illes J (ed) Neuroethics. Oxford University Press, New York

Friedson E (1973) Profession of medicine. The University of Chicago Press, Chicago

Ganley P (2002) Access to the individual: digital rights management systems and the intersection of informational and decisional privacy interests. Int J Law Inf Technol 10:241

Garde A (2007) The contribution of food labelling to the EU's obesity prevention strategy. Eur Food Feed Law Rev 6:378
Gardner J (2006) The mark of responsibility (with a postscript on accountability). In: Dowdle MW (ed) Public accountability. Cambridge University Press, New York
Garland D (2000) Ideas, institutions and situational crime prevention. In: Garland D (ed) Ethical and social perspectives on situational crime prevention. Hart Publishing, Portland, Oregon, p 1
Grabosky P, Braithwaite J (1985) Of manners gentle – enforcement strategies of Australian business regulatory agencies. Oxford University Press, Melbourne
Graham C (1998) Is there a crisis in regulatory accountability? In: Baldwin R, Scott C, Hood C (eds) A reader on regulation. Oxford University Press, New York
Harris J (2012) Chemical cognitive enhancement: is it unfair, unjust, discriminatory, or cheating for healthy adults to use smart drugs? In: Illes J, Sakakian BJ (eds) The Oxford handbook of neuroethics. Oxford University Press, New York
Hart HLA (1961) The concept of law. Oxford University Press, New York, 128
Hawkins K (1984) Environment and enforcement. Clarendon, New York
Hawkins K (2002) Law as last resort. Oxford University Press, New York
Medicines and Healthcare Devices Regulatory Authority (2006) Single use devices: implications and consequences of re-use. MHRA Device Bull DB2006(04)
Hood C, Baldwin R, Rothstein H (2001) The government of risk. Oxford University Press, Oxford, 21
Hutter B (1997) Compliance: regulation and environment, Oxford socio-legal studies. Clarendon Press, Oxford
Jelsma J (2003) Innovating for sustainability: involving users, politics and technology. Innovation 16:103
Katyal NK (2002) Architecture as crime control. Yale Law J 111:1039
Kerr I (2010) Digital locks and the automation of virtue. In: Geist M (ed) From "radical extremism" to "balanced copyright": Canadian copyright and the digital agenda. Irwin Law, Toronto
Kerr I, Bailey J (2004) The implications of digital rights management for privacy and freedom of expression. J Inf Commun Ethics Soc 2:87
Kohn LT et al (2000) To err is human: building a safer health system. National Academy Press, Washington, DC
Kwayke G, Provonost P, Makary M (2010) A call to go green in health care by reprocessing medical equipment. Acad Med 85:398
Lessig L (1999) Code and other laws of cyberspace. Basic Books, New York
Levi-Faur D (2005) The global diffusion of regulatory capitalism. Ann Am Acad Pol Soc Sci 598:12
Levy B, Spiller PT (1996) Regulators, institutions and commitment. Cambridge University Press, Cambridge
Lyons MK (2011) Deep brain stimulation: current and future clinical applications. Mayo Found 86:662
Mello M (2009) New York city's war on fat. N Engl J Med 360:2015
Mello MM, Rosenthal MB (2008) Wellness programs and lifestyle discrimination – the legal limits. N Engl J Med 359:192
Merry A, McCall Smith RA (2001) Errors, medicine and the law. Cambridge University Press, Cambridge
Ministerial Group on Nanotechnologies (2010) UK nanotechnologies strategy: small technologies, great opportunities. London, 28
Morgan B, Yeung K (2007) An introduction to law and regulation. Cambridge University Press, Cambridge, UK
BBC News (2013) Can genetically modified mosquitoes prevent disease in the US?. http://www.bbc.co.uk/news/world-us-canada-19091880

Nuffield Council on Bioethics (2002) Genetics and human behaviour: the ethical context. Nuffield Council on Bioethics, London
Oke KB et al (2013) Hybridization between genetically modified Atlantic salmon and wild brown trout reveals novel ecological interactions. Proc R Soc 280:20131047
Open Rights Group (2013) Campaign 'stop opt-out' "adult" filtering. https://www.openrightsgroup.org/campaigns/defaultblocking. Accessed 25 Nov 2013
Paul M et al (2011) Molecular pharming: future targets and aspirations. Hum Vaccin 7:375
Peckham S (2012) Slaying sacred cows: is it time to pull the plug on water fluoridation? Crit Publ Health 22:159
Raine A (2013) Dickson anatomy of violence. Allen Lane, London, pp 329–373
Reason J (2000) Human error: models and management. Brit Med J 320:768
Reidenberg JR (1998) Lex informatica: the formulation of information policy rules through technology. Texas Law Rev 76:553
Rizzo MJ, Whitman DG (2009) Little brother is watching you: new paternalism on the slippery slopes. Ariz Law Rev 51:685
Romero-Bosch A (2007) Lessons in legal history – eugenics and genetics. Mich St J Med Law 1:89
Rostain T (2010) Self-regulatory authority, markets, and the ideology of professionalism. In: Baldwin R, Lodge M, Cave M (eds) The Oxford handbook of regulation. Oxford University Press, New York
Schauer FS (1991) Playing by the rules. Clarendon, Oxford
Schlag P (2010) Nudge, choice architecture and libertarian paternalism. Mich Law Rev 108:913
Scott C (2000) Accountability in the regulatory state. J Law Soc 27:38
Sempos CT, Park YK, Barton CN, Vanderveen JE, Yetley EA (2000) Effectiveness of food fortification in the United States: the case of pellagra. Am J Publ Health 90:727
Smith AF et al (2006) Adverse events in anaesthetic practice: qualitative study of definition, discussion and reporting. Brit J Anaesth 96:715
Stier DD, Mercy JA, Kohn M (2007) Injury prevention. In: Goodman RA et al (eds) Law in public health practice. Oxford University Press, New York
Thaler R, Sunstein C (2009) Nudge. Penguin Books, London
The Independent (2013) It is a slow metabolism after all: Scientists discover obesity gene. http://www.independent.co.uk/news/science/it-is-a-slow-metabolism-after-all-scientists-discover-obesity-gene-8902235.html. Accessed 13 Nov 2013
The Roslin Institute, University of Edinburgh (2013) 'GM chickens that don't transmit bird flu'. http://www.roslin.ed.ac.uk/public-interest/gm-chickens/. Accessed 13 Nov 2013
The Rt Honourable David Cameron MP (2013) The internet and pornography: prime minister calls for action. https://www.gov.uk/government/speeches/the-internet-and-pornography-prime-minister-calls-for-action. Accessed 25 Nov 2013
Toft B (2001) External inquiry into the adverse incident that occurred at Queen's Medical Centre, Nottingham. Department of Health, London
Wadden TA, Brownell KD, Foster GD (2002) Obesity: responding to the global epidemic. J Consult Clin Psychol 70:510
White MD (2010) Behavioural law and economics: the assault on the consent, will and dignity. In: Gaus G, Favor C, Lamont J (eds) Essays on philosophy, politics & economics: integration and common research projects. Stanford University Press, Stanford, California
Winner L (1980) Do artifacts have politics? Daedalus 109:121
Yeung K (2004) Securing compliance. Hart Publishing, Oxford
Yeung K (2008) Towards an understanding of regulation by design. In: Brownsword R, Yeung K (eds) Regulating technology. Hart Publishing, Oxford
Yeung K (2011a) The regulatory state. In: Baldwin R, Cave M, Lodge M (eds) Oxford handbook on regulation. Oxford University Press, Oxford
Yeung K (2011b) Can we employ design-based regulation while avoiding brave new world. Law Innov Technol 3:1

Yeung K (2012) Nudge as fudge. Mod Law Rev 75:122
Yeung K (2016) Is design-based regulation legitimate?'. In: Brownsword R, Scotford E, Yeung K (eds) The Oxford handbook on the law and regulation of technology. Oxford University Press, Oxford, forthcoming
Yeung K, Dixon-Woods M (2010) Design-based regulation and patient safety: a regulatory studies perspective. Soc Sci Med 71:502
Zittrain J (2007) Tethered appliances, software as service, and perfect enforcement. In: Brownsword R, Yeung K (eds) Regulating technologies. Hart Publishing, Portland, Oregon

# Design for the Value of Responsibility

Jessica Nihlén Fahlquist, Neelke Doorn, and Ibo van de Poel

## Contents



J.N. Fahlquist (✉)
TU Delft / 3TU, Centre for Ethics and Technology, Delft, The Netherlands
e-mail: jessicanfahlquist@gmail.com

N. Doorn
Department of Technology, Policy and Management, TU Delft, Delft, The Netherlands
e-mail: n.doorn@tudelft.nl

I. van de Poel
Department of Ethics and Philosophy of Technology, Delft University of Technology, Delft, The Netherlands
e-mail: i.r.vandepoel@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_18

**Abstract**

The responsibility of engineers and designers for the products they design is a common topic in engineering ethics and ethics of technology. However, in this chapter we explore what designing for the value of responsibility could entail. The term "design for the value of responsibility" can be interpreted in (at least) two ways. First, it may be interpreted as a design activity that explicitly takes into account the effect of technological designs on the possibility of users (and others) to assume responsibility or to be responsible. Second, it may refer to a design activity that explicitly affects the allocation of responsibility among the ones operating or using the technology and other affected people. In this chapter, we discuss both interpretations of design for the value of responsibility. In both interpretations, a technological design can be said to affect a person's responsibility. As there are no explicit methods or approaches to guide design for responsibility, this chapter explores three cases in which design affected responsibility and develops on basis of them design heuristics for design for responsibility. These cases are the alcohol interlock for cars in Sweden, the V-chip for blocking violent television content and developmental podcasting devices in rural Zimbabwe. We conclude by raising some open issues and suggesting future work.

**Keywords**

Conditions of responsibility • Individual and collective responsibility • Distribution of responsibility • Responsibility as a virtue

## Introduction

The responsibility of engineers and designers for the products they design is a common topic in engineering ethics and ethics of technology. However, designing for the value of responsibility, which we discuss in this chapter, is an under-theorized topic. The term "design for the value of responsibility" can be interpreted in (at least) two ways. First, it may refer to a design activity that explicitly affects the possibility of users to assume responsibility or to be responsible. Second, it may interpreted as a design activity that explicitly takes into account the effect of technological designs on the allocation of responsibility among the ones operating or using the technology and other affected people. In this chapter, we discuss both interpretations of design for the value of responsibility. In both interpretations, a technological design can be said to affect a person's responsibility. As there are no explicit methods or approaches to guide design for responsibility, this chapter explores designs that do affect responsibility. Through discussing three examples of designs which *do* affect responsibility in different ways, we will tentatively propose heuristics for designing for responsibility.

Given that design decisions affect the possibilities for assuming and discharging responsibility and the allocation of responsibility, we could in principle also deliberately design for responsibility. Several authors have made the general point that

technological artifacts can delegate tasks to technologies or humans and that they affect the moral behavior of people and may shift the balance of power and control between groups (Winner 1980; Latour 1992; Verbeek 2011). All these are obviously important for the allocation of responsibility, but they are usually not discussed in such terms. There are, of course, some exceptions. Wetmore (2004), for example, discusses car design issues in relation to responsibility for safety. Some of our own work has also focused on how responsibility may be affected by technology and design (Grill and Nihlén Fahlquist 2012; Nihlén Fahlquist and Van de Poel 2012; Doorn and Van de Poel 2012; Nihlén Fahlquist in press).

The chapter is structured as follows: First, we explicate the concept of responsibility. Second, we discuss three examples where design, more or less implicitly, affected responsibility. If we have a picture of how design can and does affect responsibility, we achieve a better understanding of how design should explicitly take responsibility into account. Third, we discuss what designing for responsibility could mean. We end with discussing open issues and challenges and we draw some conclusions.

## Explication of Responsibility

The concept of responsibility contains many different notions and is used differently in different contexts (Hart 1968; Davis 2012; Van de Poel 2011). Davis distinguishes nine senses of the term (Davis 2012). Some of the notions are more relevant than others in the context of design. In this section, we will explicate those aspects of responsibility that we believe are the most relevant to design. Design can affect both individual responsibility as well as the distribution of responsibility. We discuss individual responsibility in section "Individual Responsibility" and the distribution of responsibility in section "Distribution of Responsibility".

### Individual Responsibility

We distinguish between backward-looking and forward-looking responsibility (Van de Poel 2011). Backward-looking refers to responsibility for things that happened in the past; the focus is usually or undesirable outcomes, although that is not necessary. Forward-looking responsibility refers to things not yet attained. In this context, we will understand forward-looking responsibility as responsibility as a virtue.

#### Backward-Looking Responsibility

In the traditional philosophical literature on responsibility, being *morally responsible* is usually understood as meaning that the person is an appropriate candidate for reactive attitudes, such as blame or praise (Strawson 1974; Fischer and Ravizza 1993; Miller 2004). Being morally responsible (i.e., being eligible for reactions of praise and blame) is usually taken to depend on certain conditions that have to be met before it is fair to ascribe responsibility to someone. Although academics disagree on the precise formulation, the following conditions together capture the general notion of when it is

fair to hold an agent morally responsible for (the consequences of) their actions (see Feinberg 1970; Hart and Honoré 1985; Bovens 1998; Fischer and Ravizza 1998; Corlett 2006; Doorn 2012a; Van de Poel et al. 2012):

1. Moral agency: The responsible actor is an intentional agent concerning the action. This means that the agent must have adequate possession of her mental faculties at the moment of engaging in the action. Young children and people whose mental faculties are permanently or temporarily disturbed are usually not fully held responsible for their behavior because they do not fulfill this condition. However, to put oneself knowingly and voluntarily into a situation of limited mental capacity (e.g., by drinking alcohol or taking drugs) does not, in general, exempt one from being responsible for the consequences of one's behavior. Some people phrase this condition in terms of intention, meaning that the action was guided by certain desires or beliefs.
2. Voluntariness or freedom: The action resulting in the outcome was voluntary, which means that if the actor performed the action under compulsion, external pressure or hindered by other circumstances outside the actor's control, she is not held responsible. The person must be in the position to determine his own course of action (cf. condition 1) and to act according to that.
3. Knowledge of the consequences: The actor knew, or could have known, the outcome. Ignorance due to negligence, however, does not exempt one from responsibility.
4. Causality: The action of the actor contributed causally to the outcome; in other words, there has to be a causal connection between the agent's action or inaction and the damage done.
5. Transgression of a norm: The causally contributory action was faulty, which means that the actor in some way contravened a norm.

Note that especially the first two conditions are closely interrelated. Being an intentional agent means that one has the opportunity of putting the will into effect and that one is free from external pressure or compulsion (Thompson 1980; Lewis 1991). With regard to the fifth condition, extensive debate has been going on as to what counts as a norm. In daily life the norm can be much vague than in criminal law where the norm must be explicitly formulated beforehand.

### Forward-Looking Responsibility: Responsibility as a Virtue

In this context, i.e., designing for values, we will conceive of forward-looking responsibility as responsibility as a virtue. In daily life, we talk about "responsible" people; that is, people who have certain character traits associated with a certain kind of behavior and attitudes. The word virtue means "excellence," "capacity," or "ability," and being virtuous is being able to or having the power to achieve something (Van Hooft 2006). To possess virtues is the basis of being a good person (Swanton 2005). According to Williams, a responsible person, in this sense, is able and willing to respond to a plurality of normative demands (Williams 2008). According to Van Hooft, taking responsibility includes a personal involvement and commitment.

A responsible person does not leave things to others, but feels that it is "up to me" and is willing to make sacrifices in order to get involved (Van Hooft 2006). Applied to different real-life contexts, we all more or less have an image of what a responsible person is, or more contextualized a "responsible driver" or a "responsible leader." One of us has argued that to be a responsible person requires that one is also a person who cares about fellow human beings (Nihlén Fahlquist 2010).

Because virtues are often seen as something which an agent acquires with time, upbringing, habituation, and experience, and hence can possibly be affected, we are interested in whether design can promote responsible, or irresponsible, operators and users.

## Distribution of Responsibility

We will now look at the distribution of responsibility. Responsibility can be distributed over various individuals, but it can also be distributed to collectives (collective responsibility), instead of, or in addition to the distribution of responsibility to individuals. Since collective responsibility does not exclude the attribution of responsibility to individuals within a collective, we will treat the attribution of responsibility to collectives as a special case of the distribution of (individual and collective) responsibility.

Traditionally, in the philosophical literature, individuals were seen as the sole bearers of responsibility, but as the world becomes more and more collectively organized, scholars have seen the need to assign responsibility to collectives, for example organizations and nations (May and Hoffman 1991b). The question is how the relation between individual and collective responsibility should be conceived. For example, what does collective responsibility imply for the individuals who make up that collective? If an individual's organization, or a group she is a part of, does something wrong, does this mean that she is partly responsible (e.g., May and Hoffman 1991b; May 1992; Kutz 2000; Pettit 2007; Bovens 1998)? This question requires a definition of organization, and the issue of how organized a group of people need to be in order to be assigned responsibility has also been discussed (French; May 1992). In some cases, we are dealing with the responsibility of a company or government agency; in other cases we are discussing nations (Miller 2004) or just humanity in total. In relation to the latter, consider climate change, which many people probably think should be dealt with by governments, but where individuals can contribute by everyday choices. To what extent climate change is an individual or collective responsibility has been discussed by philosophers during recent years (cf. Sinnott-Armstrong 2005; Johnson 2003; Van de Poel et al. 2012; Nihlén Fahlquist 2010). Responsibility for social problems is probably partly individual and partly collective.

In addition to responsibility ascribed to a collective of people, responsibility can also be distributed over different people. When a large number of people are involved, it may be problematic to identify the person responsible for a negative outcome. Dennis Thompson referred to this problem or situation as the "problem of

many hands" (Thompson 1980). Thompson formulated the problem in the context of the moral responsibility of public officials. Because many different officials, at various levels and in various ways, contribute to policies and the decisions of the organization, it is difficult to ascribe moral responsibility for the organization's conduct. For outsiders of an organization who want to hold someone responsible for a certain conduct, it is particularly difficult or even impossible to find any person who can be said to have independently formed and carried out a certain policy or taken some decision. The problem of many hands is now widely discussed in the literature on engineering and business ethics (Harris et al. 2005/1995; Bovens 1998; Nissenbaum 1994, 1996; Doorn and Van de Poel 2012).

Thompson's definition of the problem of many hands is rather broad and it leaves room for many different interpretations. It is therefore not surprising to see many different interpretations of this problem in the applied ethics literature. Some people see the problem of many hands primarily as the epistemic problem to identify the person responsible for harm because one does not know who actually made what contribution. This is mainly a problem for outsiders, Davis (2012) argues, because insiders generally know very well who made what contribution. This problem could therefore be avoided by making each individual's causal contribution more transparent. Other authors come with a metaphysical interpretation of the problem of many hands (Bovens 1998; Nissenbaum 1994, 1996). Since our conditions for individual responsibility do not easily generalize to collective action, we need a different conception of responsibility, these authors argue. In the philosophy literature, this track is also followed by philosophers such as Peter French (1984, 1991) and Larry May (1992), who have tried to develop special principles that hold in situations where many actors are involved. In this chapter, we hold the view that the problem of many hands refers to a situation where none of the individuals is (or can be held) responsible but in which the collective of individuals is responsible. The challenge is to avoid this problem from occurring by distributing the responsibility over different individuals such that it does not occur.

## Three Examples

In this section, we discuss three examples of a technological design that affected either individual responsibility or the distribution of responsibility, or both. Our discussion of the examples in this section is explorative. The three examples are the alcohol interlock, the V-chip for blocking violent television content, and podcasting devices in rural Zimbabwe. A description of the three cases is followed by a critical comparison.

### The Alcohol Interlock

In Sweden, according to a bill adopted by a previous government and parliament, alcohol interlocks should be made mandatory in all new cars from 2012 (Grill and Nihlén Fahlquist 2012). Although this bill has not been implemented, interlocks are

now common in coaches and taxis as a result of decisions made by individual companies. Additionally, from 2012, convicted drunk drivers are able to get their driver's license back on the condition that they drive a car which has an interlock installed (http://www.transportstyrelsen.se/sv/Vag/Alkolas/Alkolas-efter-rattfylleri/). Outside of Sweden alcohol interlocks are used in many European countries, for example, as voluntary measure by transport companies, as one part of rehabilitation programs, or in school buses (http://www.etsc.eu/documents/Drink_Driving_Monitor_July_2011.pdf).

The mandatory use of alcohol interlocks would probably decrease the problem of drunk driving. However, it would also entail a different conception of who is responsible for drunk driving. Instead of seeing it mainly as an individual responsibility[1], responsibility for drunk driving would be considered a shared responsibility between, on the one hand, the traffic "system designers" of the government and car industry and the drivers on the other.

There are some counterarguments to using this technology that are related to responsibility. The first counterargument refers to the distinction between individual and collective responsibility. In Sweden, the alcohol interlock should be seen against the background of a policy change in traffic safety, defined in the so-called Vision Zero (Swedish government 1996–1997, Nihlén Fahlquist 2006). According to Vision Zero, the system designers are ultimately responsible for traffic safety, which makes it largely a collective responsibility. System designers are defined as "those public and private organizations that are responsible for the design and maintenance of different parts of the road transport system such as roads, vehicles and transportation services as well as those responsible for different support systems for safe road traffic such as rules and regulations, education, surveillance, rescue work, care and rehabilitation" (Ibid.) Hence, system designers are primarily local and national government bodies and private companies. This is not uncontroversial, since driving has traditionally been associated with ideas of freedom of movement and autonomy, and to shift the balance and make it a societal concern may be considered undesirable. From a libertarian perspective, the alcohol interlock could be seen as reducing individual freedom and responsibility (Grill and Nihlén Fahlquist 2012; Ekelund 1999). On the other hand, most libertarians would agree that the government should protect the lives and health of individuals from the threat of other individuals. The question is where to draw the line between protection from harm and paternalism (Grill and Nihlén Fahlquist 2012; Nihlén Fahlquist 2006).

As argued in Grill and Nihlén Fahlquist (2012), the fact that collective responsibility is introduced does not necessarily mean that individual responsibility is removed. Responsibility does not have to be seen as a zero-sum game. Individuals and collectives can both be responsible for the same problem.

[1]Even in this case, agencies may be seen as responsible in the sense that they inform the public about the risks involved in drunk driving, etc., but the general idea is that the individual driver is the main responsible actor.

A second possible point of criticism is that the alcohol interlock actually deprives people of (individual) responsibility. After all, the alcohol interlock may remove the option to drive while intoxicated and as such affect the freedom/voluntariness condition. Do technologies that exclude certain behavior or persuade the user to behave in a particular way affect responsibility, and if so, is this desirable from a moral point of view? Recent works on persuasive technologies are relevant in this context. Persuasive technologies are intentionally designed to change the user's attitude, behavior, or beliefs, often by giving the user feedback of her actions (or omissions) and by trying to "suggest" to her a desired pattern of behavior (Fogg 2003; Spahn 2012). It appears that the question when behavior steering technologies are morally acceptable is somewhat analogous to the question when a technology can be considered as enhancing rather than hampering responsibility. Although a generally agreed framework for assessing how and when to use behavior steering technologies is still lacking (Spahn 2012), the fact that they are put in between the extremes of manipulation on the one hand and convincing on the other may already point to some tentative answers to the question when technologies affect responsibility in a desirable way. The alcohol interlock does not just persuade the user not to drive while intoxicated; it actually blocks the possibility of doing so. It could be argued that technologies that leave the user no choice but to behave in the "morally desirable way" are undesirable. Moral freedom – that is, the freedom to behave in either a morally desirable or undesirable way and to either praise or blame people for their behavior – is "crucial for our self-understanding as rational beings" (Yeung 2011, p. 29). If we consider the alcohol interlock to be such a technology, it may not be the most desirable technology.

## The V-Chip

The V-chip is a technological device designed to prevent children from watching violent television content. TV stations broadcast a rating as part of the program. Parents program the V-chip by setting a threshold rating and all programs above the rating are blocked by the V-chip when it is turned on. The V-chip was an added provision to President Bill Clinton's Telecommunications Act in 1996 and the device is mandatory in all television sets of 12 in. and larger, but parents can decide whether or not to use it. Interestingly, in the debate about violent television content and children, it has been argued both that the V-chip removes parental responsibility and that it facilitates parental responsibility (Nihlén Fahlquist and Van de Poel 2012).

So, how can it be that this technology is simultaneously interpreted as affecting parental responsibility negatively and positively? The first thing we have to acknowledge is that in debates about the V-chip, three different senses of responsibility are at stake. First, some arguments about the effects of the V-chip on responsibility refer to the distribution of tasks. It could be argued that the task of parents (i.e., deciding what their children watch on television) is partly taken over by program makers and rating committees that apply ratings to programs and the V-chip that blocks programs with certain ratings. In terms of tasks, there is thus a shift from parents to program makers and the rating committee.

Second, some arguments refer to responsibility as causal control over the "outcomes" of the technology at hand. In terms of control, it might be argued that program makers and the rating committee get more control over what children watch on TV. However, the parents are still in control over whether the V-chip is used and what content is blocked. Moreover, the ultimate control remains with the parents. Rather than shifting control, the V-chip seems to increase the total amount of potential control. As control could be related to the causal responsibility condition, the V-chip seems to increase the total amount of responsibility rather than diminishing parental responsibility. Third, arguments refer to parental responsibility as a virtue. Parental responsibility, conceived in this way includes two relations, i.e., the custodial relation between the parent and the child and the trustee relation between the parent and society (Bayne and Kolers 2008). The V-chip concerns both of these because the device is intended to protect children against threats to their well-being and development but also to protect society by preventing children from becoming more violent as a result of having consumed extensive media violence. Nihlén Fahlquist and Van de Poel argue that the duties involved in this case are sharable and that the non-sharable long-term parental responsibility to see to it that the more specific duties are performed is not threatened by the V-chip (Nihlén Fahlquist and Van de Poel 2012).

## Podcasting Devices in Rural Farming in Zimbabwe

The last example comes from the "ICT for Development" (ICT4D) movement and it concerns the introduction of podcasting devices in the Lower Guruve area in Zimbabwe. Information and Communication Technologies (ICT) are increasingly used in the field of development aid as a tool for empowerment (Johnstone 2007). In the past, the introduction of technical devices as development aid often failed because the devices were either not suitable for the context of developing countries or they were used in a different way than the one intended by the development organization or NGO (Oosterlaken et al. 2012).

The technological devices in this example were developed in the context of the *Local Content*, *Local Voice* project, funded by the European Commission. The Lower Guruve area in Zimbabwe is a remote, semiarid area. Most people living in this area are dependent on small-scale subsistence farming (livestock production and drought-resistant crop cultivation). The district has a low literacy rate and it lacks adequate infrastructure services; there is no electricity, running water, telephone landline, mobile phone network, or FM radio network. For economic reasons, the district does not receive appropriate agricultural training from the governmental livestock officers. In order to empower the local citizens to improve their farming practices, the NGO *Practical Action* wanted to introduce an ICT-based device that would enable people to share information while keeping the impact of the technology on the power balance in the communities to a minimum. After consultation with the local stakeholders, podcasting devices on cattle management were chosen as the most appropriate device. In addition to headphones, the podcasting devices came

with loudspeakers in order to enable collective listening while sitting under a tree in the village (Oosterlaken et al. 2012, p. 116).

This case is discussed in the development ethics literature as a successful example of how to strengthen people's capacity to support themselves, without falling into the trap of paternalism. Although the technological design may, in itself, not be innovative, the use of loudspeakers is an interesting addition from a responsibility point of view. Without the loudspeakers, individual farmers can use the podcasting devices to listen to and gain knowledge. As such, it strengthens the knowledge condition mentioned in section "Individual Responsibility." However, by listening to the podcasting devices collectively, which is facilitated by the loudspeakers, cattle management may become a collective responsibility. One of the local farmers indicated that the podcasting devices fostered "group work and group harmony that did not exist before" (quoted in Oosterlaken et al. 2012, p. 117). This indicates the technology may shift the focus from individual to collective responsibility. In this particular situation, the shift to a collective level was also considered a positive change (possibly also because the collective listening made the training on cattle management itself more effective). It is of course thinkable that for certain technologies, collective responsibility is not a desirable solution. In a hierarchical setting, for example, these devices designed for collective responsibility may not be the most appropriate ones.

## Comparison and Critical Evaluation

If we compare the three examples, we see that responsibility plays a central role in each of them. We also noticed a tension between different senses of responsibility. In the case of the alcohol interlock, for example, the option of driving drunk is removed if the device is used. The driver's freedom, in a sense, is limited. If we discuss the driver's responsibility primarily in terms of the backward-looking individual, the alcohol interlock can – at first sight – be considered to hamper rather than enhance responsibility. However, in terms of forward-looking responsibility, it could also be argued that freedom is expanded if convicted drunk drivers, who would otherwise not be allowed to drive, are allowed to drive if they have an interlock installed. And in terms of virtue, the alcohol interlock could, although this is far from necessary, possibly enhance responsibility. The alcohol interlock allows the driver to drive more responsibly or at least to prevent her from driving while intoxicated. Virtues are often seen as being developed through habituation and it is possible that the interlock gradually affects the character and attitudes of some drivers. Although one could argue that the alcohol interlock only prevents wrong behavior and does not enhance virtue (cf. Yeung 2011), voluntarily installing an alcohol interlock in one's car or installing an alcohol interlock in school busses could be seen as a sign that one cares about the effect of one's actions (or the actions of one's employees) on other human beings.

The relation between the technology and responsibility was found to be even more complex in the discussion of the V-chip. In terms of the task responsibility, it can be argued that the V-chip reduces parental freedom. However, the rating system in general provides parents with information on the basis of which they can form

their opinions about what programs their children are allowed to watch and what programs not and thus might increase control and responsibility. If interpreted in terms of virtue, parental responsibility for children's watching behavior can be considered enhanced by the V-chip. This could be the case since that responsibility arguably entails protecting one's children against violent content which potentially affects their well-being and development, and society against children who may become violent partially as a consequence of watching such content and the V-chip facilitates that kind of protection.

Both the alcohol interlock and the V-chip are devices used in developed countries. The discussion of the paternalistic effect of technologies in a developed world context is a relatively new one. It has become urgent in the context of behavior-steering technologies that deprive actors from certain possibilities. In development ethics, the risk of paternalism has been discussed for quite some time now. The list of failed development aid examples is extensive. It seems that these lessons have been taken into account in the example of the podcast devices. This example shows that adding technological options to a device instead of removing options (in this example, providing loudspeakers in addition to headphones and not replacing headphones by a loudspeaker) may avoid the problem of paternalism. This is also the reason why, from the perspective of avoiding paternalism, the V-chip is less problematic than the alcohol interlock. After all, the V-chip leaves more freedom to the end user.

One important observation that follows from the discussion of all three examples is that responsibility is not a zero-sum game. As the case of the alcohol interlock indicated, adding a collective dimension to responsibility does not necessarily mean that individuals are deprived of all responsibility. Individual and collective responsibility can coexist. This means that by clever technological design, one could increase the total amount of responsibility, just as one could decrease the total amount of responsibility by poor design.

## Designing for Responsibility

In this section we develop some tentative proposal for design for responsibility. As indicated in the introduction, there are currently no approaches for design for responsibility. On the basis of the explication of responsibility (section "Explication of Responsibility") and the examples (section "Three Examples"), we do some proposals of what a design for responsibility approach might look like.

### Design for Individual Responsibility

#### Individual Backward-Looking Responsibility

As we have seen in the examples in section "Three Examples," design can affect the conditions that have to be met before someone can be held responsible. The challenge when designing for the value of responsibility is to assess which of the conditions needs to be "improved" by the technological design, especially if a

particular condition conflicts with other conditions or with other values. If we look at the freedom condition, for example, a technological design that offers a person more options for action can be said to increase a person's freedom (and as such, strengthen this person's responsibility by way of the second condition mentioned in section "Individual Responsibility"). At the same time, the increased possibilities may also provide new possibilities for using the technological artifact in a wrongful way. Conversely, taking away possibilities for immoral behavior, as in the example of the alcohol interlock, may further responsible behavior and forward-looking responsibility as a virtue but diminish backward-looking responsibility. As such, the normative challenge of improving the responsibility conditions is not a trivial one.

Van den Hoven (1998) has argued that designers have what he calls a meta-task responsibility to see to it that the technologies they design allow users and operators to fulfill their responsibilities. In terms of the responsibility conditions this means that designers have to see to it that all the relevant responsibility conditions for users and operators are fulfilled. For example, when an operator of a chemical plant is responsible for closing down the plant under certain conditions that cause a safety hazard, the system should be so designed that the operator receives the relevant information in time and in an understandable way, and in way that it can easily be distinguished from less relevant information. This specific design requirement follows from the knowledge condition of responsibility.

In terms of the five responsibility conditions discussed in section "Explication of Responsibility," we could think of the following tentative design heuristics for individual backward-looking design for responsibility:

H1. Moral agency: Design should not diminish the moral agency of users, operators, and other stakeholders. From this view point, a moral pill that makes people behave as moral automatons without the ability to reason about what is morally desirable would be undesirable.
H2. Voluntariness or freedom: Designs should respect or improve the voluntariness of actions, e.g., by increasing the options of actions for users, operators, and other stakeholders.
H3. Knowledge: Designs should provide the right knowledge in the right form for responsibility.
H4. Causality: Designs should increase the control over outcomes of actions (with the design).
H5. Transgression of a norm: Designs should make people aware of relevant moral norms and potential transgressions of them. This can, for example, be done through feedback on the actions of users when they use a design; think of the warning sign in a car when the safety belt is not used.

These heuristics are only tentative and may be overridden in certain circumstances, especially because they may conflict with each other or with other design heuristics (derived from other relevant normative demands), relating to design for forward-looking individual responsibility or design for the distribution of responsibility as listed below.

### Individual Forward-Looking Responsibility as a Virtue

Responsibility as a virtue is least problematic if a technological artifact offers a user more options for actions, which may in turn facilitate a way out of moral dilemmas (cf. Van den Hoven et al. 2012). Washing machines that can be used in eco-mode offer the user more freedom (in the sense of, more options for action), while at the same time offering the possibility of running the household more "responsibly." In the same way, we have seen that the V-Chip may enhance responsibility as a virtue. Designing for responsibility as a virtue may become problematic when a particular technology steers our behavior in a particular direction or when the behavior encouraged by a particular technology conflicts with our ideas about behaving responsibly. In a situation of driving, we probably tend to think of behaving responsibly in terms of avoiding harm (non-maleficence). In medical practice (e.g., in a psychiatric setting), avoiding harm is known to be potentially at odds with our idea of freedom, in the sense of being free from of constraints. However, if we conceive of freedom as a possibility to do something (i.e., freedom to do things), treatment against one's will can also be considered to enhance a person's freedom. After treatment, even against the patient's will, the patient may have the capacity to do certain things she was not able to do without the treatment because her physical condition has improved thanks to, for example, medication. Similarly, one could argue that disabling a person to drive while she is, for example, drunk may in fact not have the diminishing effect on responsibility if responsibility is understood analogously, that is, as a capacity to behave responsibly.

We would propose the following tentative design heuristics for individual forward-looking design for responsibility:

H6. Behavior: Design should encourage morally desirable behavior of users. It should, however, do so in a way that respects design heuristic H1, H7, and H8.
H7. Capacity: Design should encourage the capacity of users, operators, and other stakeholders to assume responsibility as a virtue, i.e., their ability to reflect on their actions and their ability to behave responsibly.
H8. Virtue: Design should foster virtues in users, operators, and other stakeholders. As virtues are acquired character traits, design can foster them.

## Design for the Distribution of Responsibility

A main issue here is what the right balance is between individual and collective responsibility. The desirable balance may well depend on the case and the circumstances, and probably cultural differences between countries are also relevant here, as the podcasting case testifies. Still, there might be some criteria to judge the balance. One, again, is effectiveness: How effective is the struck balance between individual and collective responsibility in avoiding harm and doing good? Another criterion is moral fairness: Is the balance morally fair? Some people, for example, may consider it morally inappropriate to hold the collective responsible for negative consequences caused by individuals. Conversely, sometimes it may feel to be

morally inappropriate to single out individuals for blameworthiness rather than the collective.

In relation to distributions of responsibility, a number of criteria might be employed. As a kind of minimal condition, we might want to require that what we have called the problem of many hands will not occur. This might be understood as requiring that for each relevant issue, at least someone is responsible. In addition to such a completeness requirement, we want a distribution of responsibility to be fair. Fairness relates to the question whether or not the distribution of responsibility reflects people's intuitions of when it is justified to ascribe a certain responsibility. It is unlikely that a purely consequentialist approach is psychologically feasible. The motivational force of responsibility ascriptions that are inconsistent with basic intuitions of fairness will therefore be undermined (Kutz 2000, p. 129). These basic intuitions of fairness may also differ between people (Doorn 2012b). Finally, we often want a responsibility distribution not only to be complete and fair but also to be effective in achieving some desirable end in avoiding harm (Doorn 2012a). Conceivably, some ways of distributing responsibility are more likely to avoid harm, and to foster good than others. Completeness seems a minimal or necessary condition for effectiveness, but it is certainly not sufficient. Even if for each issue, someone is responsible, the resulting responsibility distribution is not necessarily the most effective. More generally, the criteria of completeness, fairness, and effectiveness may conflict in the sense that they single out different responsibility distributions as best. Technology may play a role in distributing responsibility in a particular way. Since a technological artifact may give the user control in varying degrees, the technology may lead to deliberate distributions of responsibility between different users, between intermediate and end users, or between producers and users. The normative challenge is to single out the relevant distribution criteria and, if there are more, how to prioritize or strike a balance between them.

On basis of the above considerations, we suggest the following design heuristics for design for the distribution of responsibility:

H9. Completeness: The design should distribute responsibility in such a way that for each relevant issue at least one individual is responsible.
H10. Fairness: The design should distribute responsibilities in a fair way over individuals.
H11. Effectiveness: The design should distribute responsibility in such a way that harm is minimized and that goods are achieved as much as possible.
H12. Cultural appropriateness: Design should strike the balance between individual and collective responsibility in a way that is culturally appropriate.

## Open Issues and Future Work

As indicated above, there is currently no methodology available for systematically designing for the value of responsibility. What we have done in this chapter is to explain the different aspects of responsibility and to identify possible challenges.

The remaining challenges and open issues are of a descriptive, normative, and engineering nature.

Descriptively, the main challenge is to describe how particular designs affect responsibility. Methods are needed for describing how design affects (1) individual responsibility (backward-looking as well as forward-looking) and (2) distributions of responsibility. With respect to individual responsibility, there is relevant work and methodology in, for example, cognitive ergonomics (relevant for the knowledge condition) and in persuasive technology (relevant for the freedom and norm condition). With respect to distributions of responsibility, the point has often be made that design affects these but most analyses are based on retrospective case studies, and there exists no methodology, as far we know, to predict such effects prospectively.

Normatively, we have identified various heuristics which need to be further developed and specified. Also on basis of the examples we gave, we see the following normative challenges here:

- How to deal with the normative tensions that follow from the different aspects of responsibility. One pertinent question is how responsible behavior (forward-looking responsibility as a virtue) can be encouraged without falling into the trap of paternalism (freedom condition for backward-looking responsibility). Another issue is how responsibility can best be distributed between individuals and collectives.
- How such normative challenges are best resolved is most likely partly dependent on context. Different resolutions must be desirable for different technologies or in different countries or cultures. So a normative framework or approach is required that is able to do justice to relevant contextual differences.

The engineering challenge is to translate the relevant descriptive and normative insights into design methodologies and engineering solutions. While finding good engineering solutions is probably to be done by designing engineers who partake in specific projects, a task which will require a good deal of creativity, the development of design methodology for design for responsibility is a more general task. We have formulated some tentative design heuristics that make a beginning with this task; however, the development of a sound design methodology would in our view first require the resolution of some of the abovementioned descriptive and normative challenges.

## Conclusion

In this chapter, we have analyzed how one could design for the value of responsibility. However, there is currently no methodology available for systematically designing for the value of responsibility. Based on an explication of different notions of responsibility and ways in which the term is used, we identified two main ways in which design can affect responsibility, i.e., (1) individual responsibility (backward-looking as well as forward-looking) and (2) the distribution of responsibility. We further

elaborated three cases, and on basis of these we developed a number of design heuristics for design for responsibility. We also identified a number of challenges for design for responsibility. These challenges are both empirical (How to design choices affect responsibility?) and normative (What is a desirable way of affecting responsibility?). It was shown that the different heuristics for design for responsibility may in some situations conflict, especially if a technological artifact limits the user's freedom. Further research is needed to develop a methodology for designing for responsibility. This may be complemented with work on behavior steering technologies and insights from cognitive economics.

## Cross-References


- ▶ Design for the Value of Regulation
- ▶ Design for the Values of Accountability and Transparency
- ▶ Design Methods in Design for Values
- ▶ Human Capabilities in Design for Values
- ▶ Mediation in Design for Values


## References


Bayne T, Kolers A (2008) Parenthood and procreation. In: Zalta EN (ed) Stanford encyclopedia of philosophy (fall 2008 edn). http://plato.stanford.edu/archives/fall2008/entries/parenthood/

Bovens M (1998) The quest for responsibility. Accountability and citizenship in complex organisations. Cambridge University Press, Cambridge

Corlett JA (2006) Responsibility and punishment. Springer, Dordrecht

Davis M (2012) "'Ain't no one here but us social forces": constructing the professional responsibility of engineers. Sci Eng Ethics 18(1):13–34

Doorn N (2012a) Responsibility ascriptions in technology development and engineering: three perspectives. Sci Eng Ethics 18(1):1–11

Doorn N (2012b) Exploring responsibility rationales in research and development (R&D). Sci Technol Hum Values 37(3):180–209

Doorn N, Van de Poel IR (2012) Editors' overview: moral responsibility in technology and engineering'. Sci Eng Ethics 18(1):69–90

Ekelund M (1999) Varning-livet kan leda till döden. En kritik av nollvisioner. Timbro, Stockholm

European Transport Safety Council (ETSC) Newsletter 14 July 2011. http://www.etsc.eu/documents/Drink_Driving_Monitor_July_2011.pdf. Accessed 16 Jan 2012

Feinberg J (1970) Doing and deserving. Essays in the theory of responsibility. Princeton University Press, Princeton

Fischer JM, Ravizza M (1993) Introduction. In: Fischer JM, Ravizza M (eds) Perspectives on moral responsibility. Cornell University Press, Ithaca, pp 1–41

Fischer JM, Ravizza M (1998) Responsibility and control. A theory of moral responsibility. Cambridge University Press, Cambridge

Fogg BJ (2003) Persuasive technology: using computers to change what we think and do. Morgan Kaufmann, Amsterdam/Boston

French PA (1984) Collective and corporate responsibility. Columbia University Press, New York

French PA (1991) The corporation as a moral person. In: May L, Hoffman S (eds) Collective responsibility: five decades of debate in theoretical and applied ethics. Rowman & Littlefield, Savage

Grill K, Nihlén Fahlquist J (2012) Responsibility, paternalism and alcohol interlocks. Publ Health Ethics 5(2):116–127
Harris CE, Pritchard MS, Rabins MJ (2005/1995) Engineering ethics: concepts and cases, 3rd edn. Wadsworth, Belmont
Hart HLA (1968) Punishment and responsibility. Essays in the philosophy of law. Clarendon, Oxford
Hart HLA, Honoré T (1985) Causation in the law. Clarendon, London
Johnson BL (2003) Ethical Obligations in a Tragedy of the Commons. Environmental Values, 12(3):271–287
Johnstone J (2007) Technology as empowerment: a capability approach to computer ethics. Eth Inf Technol 9(1):73–87
Kutz C (2000) Complicity: ethics and law for a collective age. Cambridge University Press, New York
Latour B (1992) Where are the missing masses? In: Bijker W, Law J (eds) Shaping technology/building society; studies in sociotechnical change. MIT Press, Cambridge, MA, pp 225–258
Lewis HD (1991) Collective responsibility. In: May L, Hoffman S (eds) Collective responsibility: five decades of debate in theoretical and applied ethics. Rowman & Littlefield, Savage
May L (1992) Sharing responsibility. University of Chicago Press, Chicago
May L, Hoffman S (1991a) Introduction. In: May L, Hoffman S (eds) Collective responsibility: five decades of debate in theoretical and applied ethics. Rowman & Littlefield, Savage
May L, Hoffman S (eds) (1991b) Collective responsibility: five decades of debate in theoretical and applied ethics. Rowman and Littlefield, Savage
Miller D (2004) Holding nations responsible. Ethics 114:240–268
Nihlén Fahlquist J (2006) Responsibility ascriptions and vision zero. Accid Anal Prev 38(6):1113–1118
Nihlén Fahlquist J (2010) The problem of many hands and responsibility as the virtue of care. managing in critical Times – philosophical responses to organisational turbulence proceedings. St Anne's College, Oxford. 23–26 July 2009
Nihlén Fahlquist J (2013) Responsibility and privacy – ethical aspects of using GPS to track children. Child Soc (in press)
Nihlén FJ, Van de Poel IR (2012) Technology and parental responsibility – the case of the V-chip. Sci Eng Ethics 18(2):285–300
Nissenbaum H (1994) Computing and accountability. Commun ACM 37(1):73–80
Nissenbaum H (1996) Accountability in a computerized society. Sci Eng Ethics 2(1):25–42
Oosterlaken ET, Grimshaw D, Janssen P (2012) Marrying the capability approach with appropriate technology and STS – the case of podcasting devices in Zimbabwe. In: Oosterlaken ET, van den Hoven MJ (eds) The capability approach, technology and design. Springer, Dordrecht
Pettit P (2007) Responsibility Incorporated. Ethics 117: 171–201
Sinnott-Armstrong W, Howarth RB (eds) (2005) Perspectives on Climate Change: Sci Econ Polit Ethics. Elsevier, Amsterdam
Spahn A (2012) And lead us (Not) into persuasion. . .? Persuasive technology and the ethics of communication. Sci Eng Ethics 18(4):633–650
Strawson PF (1974) Freedom and resentment. In: Strawson PF (ed) Freedom and resentment and other essays. Methuen, London, pp 1–25
Swanton, C. 2005. Virtue Ethics. A Pluralistic View. Oxford University Press: Oxford
Swedish Government (1997) Nollvisionen och det trafiksäkra samhället, Regeringsproposition, 1996–1997, vol 137
Thompson DF (1980) Moral responsibility and public officials. Am Political Sci Rev 74:905–916
Van de Poel IR (2011) The relation between forward-looking and backward-looking responsibility. In: Vincent N, Van de Poel I, Van den Hoven J (eds) Moral responsibility. Beyond free will and determinism. Springer, Dordrecht, pp 37–52
Van de Poel IR, Nihlén Fahlquist J, Doorn N, Zwart SD, Royakkers LMM (2012) The problem of many hands: climate change as an example. Sci Eng Ethics 18(1):49–67

Van den Hoven MJ (1998) Moral responsibility, public office and information technology. In: Snellen ITM, Van de Donk WBHJ (eds) Public administration in an information age. A handbook. Ios Press, Amsterdam, pp 97–111
Van den Hoven MJ, Lokhorst GJ, Van de Poel IR (2012) Engineering and the problem of moral overload. Sci Eng Ethics 18(1):143–155
Van Hooft, S. 2006. Understanding Virtue Ethics. Acumen: Chesham
Verbeek P-P (2011) Moralizing technology: understanding and designing the morality of things. The University of Chicago Press, Chicago/London
Wetmore JM (2004) Redefining risks and redistributing responsibilities: building networks to increase automobile safety. Sci Technol Hum Values 29(3):377–405
Williams, G. 2008. Responsibility as a Virtue. Ethical Theory and Moral Practice 11(4):455–470
Winner L (1980) Do artifacts have politics? Daedalus 109:121–136
Yeung K (2011) Can we employ design-based regulation while avoiding brave New world? Law Innov Technol 3(1):1–29

# Design for the Value of Safety

Neelke Doorn and Sven Ove Hansson

## Contents



**Abstract**

Two major methods for achieving safety in engineering design are compared: safety engineering and probabilistic risk analysis. Safety engineering employs simple design principles or rules of thumb such as inherent safety, multiple barriers, and numerical safety margins to reduce the risk of accidents. Probabilistic risk analysis combines the probabilities of individual events in event chains leading to

N. Doorn (✉)
Department of Technology, Policy and Management, TU Delft, Delft, The Netherlands
e-mail: n.doorn@tudelft.nl

S.O. Hansson
Division of Philosophy, Royal Institute of Technology, Stockholm, Sweden
e-mail: soh@kth.se



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_19

accidents in order to identify design elements in need of improvement and often also to optimize the use of resources. It is proposed that the two methodologies should be seen as complementary rather than as competitors. Probabilistic risk analysis is at its advantage when meaningful probability estimates are available for most of the major events that may contribute to an accident. Safety engineering principles are more suitable to deal with uncertainties that defy quantification. In many design tasks, the combined use of both methodologies is preferable.

**Keywords**
Design • Risk • Probabilistic risk analysis • Safety factor • Uncertainty, Safety engineering

## Introduction

Enhancing safety and avoiding or mitigating risks have been a central concern of engineering as long as there have been engineers. Already in the earliest engineering codes, it was established that engineers should hold paramount the safety of the general public (Davis 2001). Following the definition of design as an "activity in which certain functions are translated into a blueprint for an artifact, system, or service that can fulfill these functions" (Van de Poel and Royakkers 2011, p. 166), a distinction is usually made between functional and nonfunctional requirements, the latter referring to requirements that have to be met but that are not necessary for the artifact, system, or service to fulfill its intended function. Contrary to most of the other values discussed in this volume, the value of safety is almost always conceived as a ubiquitous though often implicit functional requirement. Even if it is not stated explicitly in the design requirements, the need to make the design "safe" is almost always presupposed. The importance that is assigned to safety will differ, though, and there are different ways to take safety into account during design. In this chapter, we will discuss two main approaches to designing for the value of safety: safety engineering and probabilistic risk analysis. We first define the key terms, also in relation to the different engineering domains (section "Definitions"). After that, we present the two main approaches (section "Current Approaches"), followed by a discussion of the pros and cons of both approaches (section "Discussion of the Two Approaches"). The approaches are illustrated with two examples from civil engineering (section "Experiences and Examples"), followed by a critical evaluation, including some open issues (section "Critical Evaluation"). In the concluding section "Conclusions," we summarize the findings.

## Definitions

Technological risk and safety is an area in which the terminology is far from well established. The definition of key terms not only differs between disciplines and contexts (such as engineering, natural sciences, social sciences, and public

discussion), it often differs between different branches and traditions of engineering as well. These differences depend largely on lack of communication between different expert communities, but there is also a normative or ideological element in the terminological confusion. Different uses of "risk" and "safety" tend to correlate with different views on how society should cope with technological risk.

## Risk

To start with the notion of risk, it is important to distinguish between *risk* and *uncertainty*. This distinction dates back to work in the early twentieth century by the economists J. M. Keynes and F. H. Knight (Keynes 1921; Knight 1935[1921]). Knight pointed out that "[t]he term 'risk', as loosely used in everyday speech and in economic discussion, really covers two things which, functionally at least, in their causal relations to the phenomena of economic organization, are categorically different." In some cases, "risk" means "a quantity susceptible of measurement," while in other cases "something distinctly not of this character." He proposed to reserve the term "uncertainty" for cases of the non-quantifiable type and the term "risk" for the quantifiable cases (Knight 1935[1921], pp. 19–20).

This terminological reform has spread to other disciplines, including engineering, and it is now commonly assumed in most scientific and engineering contexts that "risk" refers to something that can be assigned a probability, whereas "uncertainty" may be difficult or impossible to quantify.

In engineering, risk is quantified in at least two different ways. The first refers to the probability of an unwanted event which may or may not occur (cf. the quote, "the risk of a melt-down during this reactor's life-time is less than one in 10,000"). The second conception of risk refers to the statistical expectation value of unwanted events which may or may not occur. Expectation value means probability-weighted value. Hence, if for the construction of some large infrastructural project the probability of death is 0.005 % for each year worked by an individual and the total construction work requires 200 person-years of work, then the expected number of fatalities from this operation is $200 \times 0.00005 = 0.01$. The risk of fatalities in this operation can then be said to be 0.01 deaths. Expectation values have the important property of being additive. Suppose that a certain operation is associated with a 1 % probability of an accident that will kill five persons and also with a 2 % probability of another type of accident that will kill one person. Then the total expectation value is $0.01 \times 5 + 0.02 \times 1 = 0.07$ deaths. In similar fashion, the expected number of deaths from a hydraulic dam is equal to the sum of the expectation values for each of the various types of accidents that can occur in or at the dam.

It should be noted, however, that in everyday language, "risk" is often used without reference to probability. Furthermore, although uncertainty and risk are commonly defined as two mutually exclusive concepts, it is common practice to use "uncertainty" in lieu of "risk or uncertainty." Then "uncertainty" is used as a general term for lack of knowledge (whether probabilistic or not), and risk is a

special form of uncertainty, characterized by the availability of a meaningful probability estimate. In what follows, we will adhere to this practice and use "uncertainty" in the broad sense that covers (probabilizable) risk.

Even in cases when the plausibility of a danger can be meaningfully summarized in a probability estimate, there may yet remain significant uncertainties about the accuracy of that estimate. In fact, only very rarely are probabilities known with certainty. Even if we have extensive knowledge of the design of a nuclear power plant, for example, we do not know the exact probability of failure of the plant. The probability that a tsunami would cause a meltdown of the Fukuyama nuclear reactors in Japan, as it did in 2011, could not have been predicted with any accuracy beforehand. Therefore, even if a decision problem is treated as a decision "under risk," this does not mean that the decision in question is made under conditions of completely known probabilities. Rather, it means that a choice has been made to simplify the description of this decision problem by treating it as a case of known probabilities. This is practically important in engineering design. Some of the probability estimates used in risk calculations are quite uncertain. Such *uncertainty about probabilities* should be taken into account when probabilistic analyses are used for decision-guiding purposes.

## Safety

The concept of safety is sometimes used in an absolute, sometimes in a relative, sense. In order to illustrate the meaning of absolute safety, suppose that you buy a jacket that is promised to be of fireproof fabric. Later, it actually catches fire. Then you might argue, if you apply the absolute notion of safety, that you were not safe from fire in the first place. If the producer of the jacket tries to argue that you were in fact safe since the fabric was highly unlikely to catch fire, you would probably say that he was simply wrong. In some contexts, therefore, "I am safe against the unwanted event X" is taken to mean that there is no risk at all that X will happen.

Technical safety has often been defined as absolute safety. For example, in research on aviation safety, it has been claimed that "Safety is by definition the absence of accidents" (Tench 1985). However, in practice absolute safety is seldom achievable. For most purposes, it is therefore not a very useful concept. Indeed, the US Supreme Court has supported a non-absolute interpretation, stating that "safe is not the equivalent of 'risk free'" (Miller 1988, p. 54). With this interpretation, a statement such as "this building is fire-safe" can be read as a short form of the more precise statement: "The safety of this building with regard to fire is as high as can be expected in terms of reasonable costs of preventive actions." In this vein, the US Department of Defense has stated that safety is "the conservation of human life and its effectiveness, and the prevention of damage to items, consistent with mission requirements" (Miller 1988, p. 54).

Usage of the term "safe" (and derivatives such as "safety") in technical applications, e.g., in aviation safety, highway safety, etc., vacillates between the absolute concept ("safety means no harm") and a relative concept that only requires the risk

reduction that is considered feasible and reasonable. It is not possible to eliminate either of these usages, but it is possible to keep track of them and avoid confusing them with each other.

Safety is usually taken to be the inverse of risk: when the risk is high, then safety is low, and conversely. This may seem self-evident, but the relationship between the two concepts is complicated by the fact that as we saw in Subsection "Risk," the concept of risk is in itself far from clear. It has been argued that if risk is taken in the technical sense as statistical expectation value (expected harm), then safety cannot be the antinomy of risk, since other factors such as uncertainty have to be taken into account when assessing safety (Möller et al. 2006). With a broader definition of risk, an antonymic relationship between the two concepts may be more plausible.

## Terminological Differences Between Engineering Domains

Safety engineering has "separate origins" in many different engineering disciplines. Due in part to lack of communication between these disciplines, in part to differences in their technical tasks and social conditions, these disciplines have developed different approaches to safety. This is also reflected in their terminological usages. To illustrate these differences, let us compare three widely different engineering disciplines: nuclear engineering, civil engineering, and software engineering.

Beginning with the concept of risk, nuclear engineering represents an extreme case among the engineering disciplines. Nuclear engineers have pioneered the use of probabilistic analysis in risk management. In their industry, by a "risk-based approach" is meant the use of probabilities to characterize hazards and prioritize their abatement. However, non-probabilistic thinking also has a place in the nuclear industry. In so-called "deterministic" analysis of nuclear risks, the focus is on what can at all happen under unfavorable circumstances.

Software engineering represents the other extreme. Software developers spend much of their work time trying to avoid various undesirable events such as errors due to unusual or unforeseen inputs, intrusion by hackers, and operator mistakes caused by confusing human-machine interfaces. However, only seldom do they calculate or estimate the probabilities of these possible errors. In the vast majority of cases, they treat risks in a way that nuclear engineers would call "deterministic." Thus, when talking about a "risk," they refer to an undesired event or event chain rather than a probability or expectation value.

In civil engineering, the standard approach is non-probabilistic. The tolerance of structural failures, such as a collapsing house, bridge, or dam, is very low. Design and construction work takes place under the assumption that such events should be avoided at almost any price. Traditionally, numerical probabilities of such failures have not been calculated as part of routine construction work. Instead, less complicated rules of thumb, including numerical safety factors, have been used to obtain the desired low probabilities. However, recently probabilistic analysis has increasingly been applied, particularly in large constructions.

The concept of safety also has different connotations in the three areas. In the nuclear industry, "safety" usually refers to the avoidance of large accidents that would pose a risk to both workers and the public. In building and construction, "safety" usually refers to worker safety. This difference is quite appropriate; large accidents that put the public at risk are a much more serious problem in nuclear engineering than in most civil engineering projects, whereas the building industry in most countries has a much worse record of workplace accidents than the nuclear industry. In software engineering, safety is less often referred to; instead "security" is the common antonym of "risk" in this area.

## Current Approaches

In this section we discuss two main approaches to design for safety, viz., safety engineering and probabilistic risk analysis. Safety engineering is the older of the two, possibly going as far back as the earliest use of technological artifacts. The second approach, probabilistic risk analysis, is of more recent date and has been developed from the late 1960s onwards.

### Safety Engineering

With the development of technological science, safety engineering has gained recognition as an academic discipline, and various attempts have been made to systematize its practices. Since the discussion of safety engineering is fragmented over different disciplines, there is no unified way to do this. However, the following three principles of safety engineering summarize much of its fundamental ideas:

1. *Inherently safe design*. A recommended first step in safety engineering is to minimize the inherent dangers in the process as far as possible. This means that potential hazards are excluded rather than just enclosed or otherwise coped with. Hence, dangerous substances or reactions are replaced by less dangerous ones, and this is preferred to using the dangerous substances in an encapsulated process. Fireproof materials are used instead of inflammable ones, and this is considered superior to using flammable materials but keeping temperatures low. For similar reasons, performing a reaction at low temperature and pressure is considered superior to performing it at high temperature and pressure in a vessel constructed for these conditions (Hansson 2010).
2. *Safety factors*. Constructions should be strong enough to resist loads and disturbances exceeding those that are intended. A common way to obtain such safety reserves is to employ explicitly chosen numerical safety factors. Hence, if a safety factor of 2 is employed when building a bridge, then the bridge is calculated to resist twice the maximal load to which it will in practice be exposed (Clausen et al. 2006).

3. *Multiple independent safety barriers*. Safety barriers are arranged in chains. The aim is to make each barrier independent of its predecessors so that if the first fails, then the second is still intact, etc. Typically the first barriers are measures to prevent an accident, after which follow barriers that limit its consequences, and finally rescue services as the last resort.

In the remainder of this subsection, we will focus on safety factors, being one of the most widely applied principles of safety engineering. It is generally agreed in the literature on civil engineering that safety factors are intended to compensate for five major types of sources of failure:

(1) Higher loads than those foreseen
(2) Worse properties of the material than foreseen
(3) Imperfect theory of the failure mechanism in question
(4) Possibly unknown failure mechanisms
(5) Human error (e.g., in design) (Knoll 1976; Moses 1997)

The first two of these can in general be classified as variabilities, that is, they refer to the variability of empirical indicators of the propensity for failure. They are therefore accessible to probabilistic assessment (although these assessments may be more or less uncertain). In the technical terminology that distinguishes between risk and uncertainty, they can be subsumed under the category of risk. The last three failure types refer to eventualities that are difficult or impossible to represent in probabilistic terms and therefore belong to the category of (non-probabilizable) uncertainty.

In order to provide adequate protection, a system of safety factors will have to consider all the integrity-threatening mechanisms that can occur. For instance, one safety factor may be required for resistance to plastic deformation and another one for fatigue resistance. Different loading situations may also have to be taken into account, such as permanent load ("dead load," i.e., the weight of the building) and variable load ("live load," i.e., the loads produced by the use and occupancy of the building), the safety factor of the latter being higher because of higher variabilities. Similarly, components with widely varying material properties (e.g., brittle materials such as glass) are subject to higher safety factors than components of less variable materials (e.g., steel and metallic materials). Geographic properties may be taken into account by applying additional wind and earthquake factors. Design criteria employing safety factors can be found in numerous building codes and other engineering standards.

## Probabilistic Risk Analysis

In the late 1960s, rapidly growing public opposition to new technologies gave rise to a new market for applied science: a market for experts on risks and on the

public's attitudes to risks. The demand came mostly from companies and institutions associated with the technologies that had been subject to public opposition. The supply was met by professionals and academics with training in the natural, behavioral, and social sciences. Most of their undertakings focused on chemicals and on nuclear technology, the same sources of risk that public opposition had targeted on. The new field was institutionalized as the discipline of probabilistic risk analysis, with professional societies, research institutes, and journals of its own. From the beginning, calculations of probabilities had a central role in the new discipline. In engineering, the terms probabilistic risk analysis and probabilistic risk assessment (often abbreviated to PRA) are mostly used interchangeably. We will use the term probabilistic risk analysis to refer to the approach of using probabilistic estimates and PRA to refer to the probabilistic evaluation of a particular design or artifact. The term probabilistic design will be used to refer to design methods that are based on probabilistic risk analysis.

Probabilistic risk analysis has largely been developed in the nuclear industry. Although the engineers designing nuclear reactors in the 1950s and 1960s aimed at keeping the probability of accidents very low, they lacked means to estimate these probabilities. In the late 1960s and early 1970s, methodology was developed to make such estimates. The first comprehensive PRA of a nuclear reactor was the Rasmussen report (WASH-1400) that was published in 1975 (Rasmussen 1975; Michal 2000). Its basic methodology is still used, with various improvements, both in the nuclear industry and in an increasing number of other industries as a means to calculate and efficiently reduce the probability of accidents.

Key concepts in probabilistic risk analysis are failure mode and effect analysis (FMEA) and fault trees. FMEA is a systematic approach for identifying potential failure modes within a system. The potential modes of failure of each component of the system are investigated, and so are the ways in which these failures can propagate through the system (Dhillon 1997). A failure mode can be any error or defect in the design, use, or maintenance of a component or process in the system. In effect analysis the consequences of those failures are investigated. The next step is to identify for each of these failure modes the accident sequences that may lead to its occurrence. Typically, several such sequences will be identified for each event. Each sequence is a chain containing events such as mechanical equipment failure, software failure, lacking or faulty maintenance, mistakes in the control room, etc. Next, the probability of each of these accident sequences is calculated, based on the probability of each event in the sequence. Some of these probabilities can be based on empirical evidence, but others have to be based on expert estimates. The final step in a probabilistic risk analysis consists in combining all this information into an overall assessment. It is common to combine the different failure modes into a so-called fault tree. Fault trees depict the logical relations between the events leading to the ultimate failure. Usually, these relations are limited to AND-gates and OR-gates, indicating whether two events are both necessary for failure (AND-gate) or each of them separately leads to failure (OR-gate) (Ale 2009).

Figure 1 shows a (simplified) fault tree for a train accident. A train accident occurs if any of the following three events occur: fire, collision, or derailment

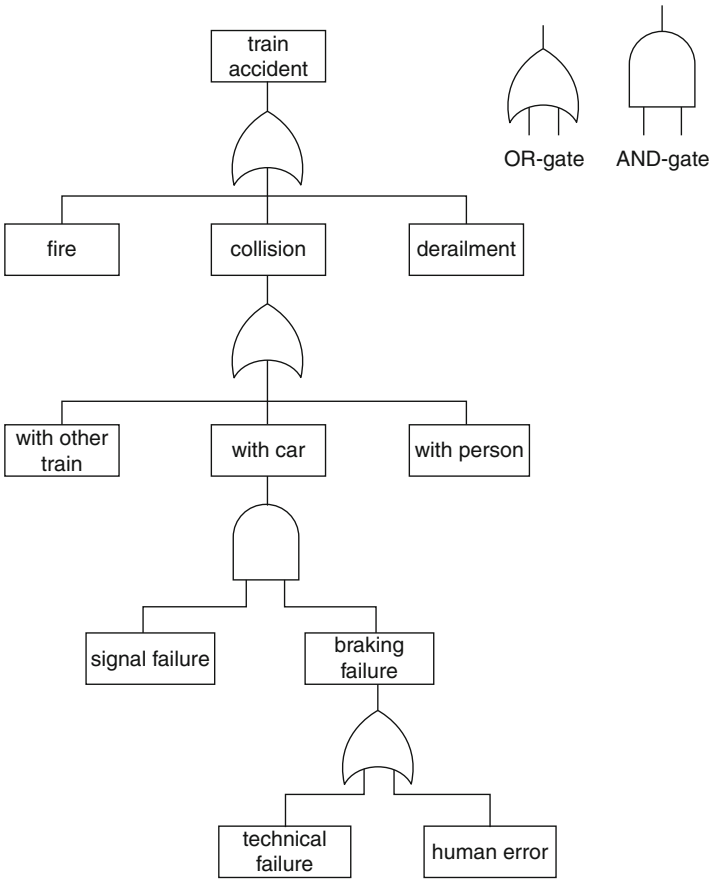



**Fig. 1** Fault tree of train accident

(OR-gate). A collision with a car occurs if both the signals and the braking fail (AND-gate).[1] Braking can fail either due to technical failure or due to human error (OR-gate).

In the early days of probabilistic risk analysis, the overall assessment often included a total probability of a major accident and/or a statistical expectation value for the number of deaths per year resulting from accidents in the plant. Today, most PRA specialists in the nuclear industry consider such overall calculations to be too uncertain. Instead, their focus is on using analyses of

[1]In this simplified example, it is assumed that in case of properly functioning signals, the driver will also stop at the halt line. Hence, for a collision to occur, it is both necessary that the signals fail and that the driver is not able to brake in time.

accident sequences to identify weaknesses in the safety system. According to one leading expert, the final step in a PRA

> . . . is to rank the accident sequences according to their probability of occurrence. This is done because risk must be managed; knowing the major contributors to each undesirable event that was defined in the first step is a major element of risk management. Also ranked are the SSCs – systems, structures, and components – according to their contribution to the undesirable event. (Michal 2000, pp. 27–28)

The same basic methodology can be used in civil engineering. In the early 2000s, the Joint Committee on Structural Safety (JCSS) developed a Probabilistic Model Code for full probabilistic design. The code was intended as the operational part of national and transnational building codes that allow for probabilistic design but do not give any detailed guidance (Vrouwenvelder 2002). Contrary to nuclear engineering, civil engineering uses probabilistic risk analysis more to dimension individual components than to identify and analyze full accident sequences (JCSS 2001; Melchers 2002). This difference depends in part on the complicated redistribution of the load effects after each component failure, which makes it difficult to predict the behavior of the system as a whole (Ditlevsen and Madsen 2007 [1996]). However, attempts are made to broaden the scope of probabilistic risk analysis to infrastructure systems as a whole rather than single construction elements in such systems (Blockley and Godfrey 2000; Melchers 2007).

## Discussion of the Two Approaches

In the literature, several arguments have been given for and against the replacement of traditional safety engineering by probabilistic risk analysis.

### Arguments for Using Probabilistic Risk Analysis in Design

Two major arguments have been proposed in support of design methods based on probabilistic risk analysis, viz., economic optimization and fitness for policy making.

#### Economic Optimization

The first, and probably most important, argument in favor of probabilistic methods is that their output can be used as an input into economic optimization. Some argue that economic optimization of risk management measures is in fact the main objective of probabilistic risk analysis (Guikema and Paté-Cornell 2002). Traditional approaches in safety engineering, such as safety factors, provide regulatory bounds that may sometimes be overly conservative (Chapman et al. 1998). There is, for instance, no way to translate the difference between using the safety factor 2.0 and the safety factor 3.0 in the design of a bridge into a quantifiable effect on safety. Without a quantifiable effect (such as reduction in the expected number of fatalities), it is impossible to calculate the marginal cost of risk reduction, and therefore

economic optimization of design is not possible. In contrast, a PRA that provides accident probabilities as outcomes makes it possible to calculate the expected gains from a safer design. This is what is needed for an optimization of the trade-off between risks and benefits (Paté-Cornell 1996; Moses 1997).

Such optimization may involve trade-offs against other factors than money. A risk can, for instance, be weighed against other risks that countermeasure against the first risk brought about (Graham and Wiener 1995). It is also common for overdesign to have a price in terms of excess usage of energy and other natural resources. Accident probabilities obtained in a PRA can be used as inputs into a risk-benefit analysis (RBA) or cost-benefit analysis (CBA) in which different types of advantages and disadvantages are taken into account (Rackwitz 2004).

The major problem with this argument for probabilistic risk analysis is that the outputs of PRAs are not always accurate enough to be used as inputs into economic analysis. Some relatively small and standardized infrastructure projects have effects that can be described fairly accurately in probabilistic terms. This applies, for instance, to some safety measures in road traffic such as central barriers on highways (Mak et al. 1998) or pedestrian crosswalks at intersections (Zegeer et al. 2006), for which the expected number of saved lives can be estimated with reasonable accuracy and weighed against the economic costs. In larger and more complex projects, the probabilistic quantifications of the effects of safety measures are generally not considered accurate enough to be used as direct inputs into economic analysis. For example, the safety of a gravity dam, a hydraulic structure that is supposed to be stable by its own weight, is largely dependent on seismic activity and on how the structure responds to it. Both can at most be quantified roughly, making it difficult to provide accurate accident probabilities (Abbas and Manohar 2002). In cases like this, it is therefore recommended to develop a robust structural design rather than an economically optimized one (Takewaki 2005). Similar problems are faced in the design of other large infrastructure projects, such as flood defense structures and offshore facilities. In summary, the argument that probabilistic risk analysis provides means for economic optimization is not valid for probabilistic risk analysis in general but only for those probabilistic risk analyses that provide probability estimates that are well calibrated with actual frequencies.

### Fitness for Policy Making

A second advantage of probabilistic approaches concerns the organizational separation between risk assessment and risk management. In the 1970s the unclear role of scientists taking part in risk policy decisions led to increasing awareness of the distinction between scientific assessments and policy decisions based on these assessments. This resulted in what is now the standard view on the risk decision process, according to which its scientific and policy-making parts should be strictly distinguished and separated. This view was expressed in a 1983 report by the US National Academy of Sciences (National Research Council 1983). The decision procedure is divided into two distinct parts to be performed consecutively. The first of these, commonly called *risk assessment*, is a scientific undertaking. It consists of

collecting and assessing the relevant information and using it to characterize the nature and magnitude of the risk. The second procedure is called *risk management*. Contrary to risk assessment, it is not a scientific undertaking. Its starting point is the outcome of risk assessment, which it combines with economic and technological information pertaining to various ways of reducing or eliminating the risk and also with political and social information. Its outcome is a decision on what measures – if any – should be taken to reduce the risk. In order to protect risk assessments from being manipulated to meet predetermined policy objectives, it was proposed to separate risk assessment organizationally from risk management. Compared to the safety engineering approach, probabilistic risk analysis seems more compatible with this organizational division between risk assessment and risk management. The selection of safety margins and other engineering measures to enhance safety is a value-dependent exercise, but it tends to be difficult to separate from scientific and technological considerations. In contrast, a PRA can be performed on the basis of scientific information alone. It is then up to the decision makers to set the acceptable probability of failure.

However, in most fields of engineering, there is in practice no separation between risk assessment and risk management. Technical standards, exposure limits, etc. are typically set by groups of experts who are entrusted both with assessing the scientific data and proposing regulation (Hansson 1998). In structural engineering, for example, the establishment of the European construction standard (Eurocodes) was characterized by organizational integration of risk assessment and risk management (Clausen and Hansson 2007). Similarly, in hydraulic engineering, Vrijling et al. (1998) developed a unified framework assessing safety in terms of acceptable individual and societal risks levels, which they derived from accident statistics and a postulated value of human life. Although the authors admit that the final judgment is political, the proposed approach merges risk assessment and management into one decision procedure.

These examples illustrate how the notions of probability and probabilistic design enter the domain of risk management where decisions on the acceptance of risks are made. Although probabilistic risk analysis in principle facilitates a clear distinction between risk assessment and risk management, the acceptable risk levels in a PRA are often decided in the community of safety experts who make the assessment as well. Hence, the actual organizational structure does not support or encourage a separation between risk assessment and risk management. This is a severe limitation on the practical applicability of the proclaimed advantage of probabilistic risk analysis that it is well suited for making this separation.

## Arguments for the Safety Engineering Approach

In this section, we discuss the four arguments that we have found in the literature in favor of safety engineering approaches such as safety factors rather than probabilistic risk assessment. These arguments refer to computational costs, simplicity, residual uncertainties, and security.

### Computational Costs

Probabilistic models promise to provide accurate estimates of failure probabilities that depend on many different input variables. The costs for data acquisition and computation tend to increase rapidly with the number of input variables. In practice, this leads either to unworkably long time for the analysis or to simplifications of the model that unavoidably lead to a decrease in accuracy. Especially when the additional time also involves delays in the design and engineering process itself, the simplicity of the safety factor approach may be an advantage, also from a cost-benefit point of view. In the building industry, the efficiency of the building process is often more important for cost-efficiency than the amount of material used. Hence, reducing the construction time may be economically preferable to saving construction material.

### Simplicity

The simplicity of the safety engineering approach can make mistakes less likely. The importance of simplicity in safety work is known from chemical plant design. Plants with inherently safer technologies tend to be simpler in design, easier to operate, and more error tolerant (Overton and King 2006). Similarly, simpler calculation or design methods may be preferable to complex ones since they reduce the likelihood of mistakes in the calculations and, hence, the likelihood of mistakes in the construction itself.

### Residual Uncertainties

One of the disadvantages of probabilistic design methods is that they can take potential adverse effects into account only to the extent that their probabilities can be quantified (Knoll 1976; Clausen et al. 2006; Hansson 2009a). Although attempts are made to quantify as many elements as possible, including human errors, this can at most be done approximately. In practice, these difficulties may lead to a one-sided focus on those dangers that can be assigned meaningful probability estimates. Probabilistic approaches tend to neglect potential events for which probabilities cannot be obtained (Knoll 1976; Hansson 1989). Safety factors, to the contrary, are intended to compensate also for in practice unquantifiable uncertainties such as the possibility that there may be unknown failure mechanisms or errors in one's own calculations. It is a rational and not uncommon practice to set a higher safety factor to compensate for uncertainty. This is done routinely in toxicology (Santillo et al. 1998; Fairbrother 2002), and it seems sensible to do so in other fields as well.

The use of safety factors is not the only method in safety engineering that takes uncertainties into account. The same applies to the other safety principles mentioned in section "Safety Engineering," namely, inherent safety and multiple safety barriers. These and other safety engineering principles introduce some degree of redundancy in the system, which is often an efficient way to protect also against dangers for which meaningful probability estimates are unavailable. Such "extra" safety may not be defensible from a cost-benefit perspective, but it may nevertheless be justified from the perspective of protection against uncertainties

(e.g., uncertainties about the probabilities of known risks and about unknown failure modes). For an example of this, suppose that a ship builder comes up with a convincing plan for an unsinkable boat. A PRA shows that the probability of the ship sinking is incredibly low and that the expected cost per life saved by lifeboats would be exceptionally high. There are several reasons why the ship should still have lifeboats: the calculations may possibly be wrong, some failure mechanism may have been missed, or the ship may be exposed to some unknown danger. Although the PRA indicates that such measures are inefficient, we cannot trust the PRA to be certain enough to justify a decision to exclude lifeboats from the design. Similar arguments can be used, for instance, for introducing an extra safety barrier in a nuclear reactor, although a PRA indicates that it is not necessary. This is, of course, not an argument against performing PRAs but an argument against treating their outcomes as the last word on what safety requires.

### Security and Vulnerability

A fourth argument in favor of the safety factor approach is related to security threats. So far, we have focused on safety, that is, the protection against unintended harm. However, the attacks on the New York Twin Towers on September 11, 2001, showed that not only "acts of nature" threaten the integrity of engineering structures. We also need protection against another type of threats, namely, those following from intended harm. This distinction is often expressed with the terms safety (against unintended harm) and security (against intended harm). Golany et al. (2009) refer to the former as probabilistic risk and the latter as strategic risk (where "strategic" refers to environments in which intentional actions are taken; it should be noted that Golany et al. do not discuss the epistemic uncertainties that may also be present in strategic situations). An important distinction is that in the latter case, there is an adversary who is capable of intelligent behavior and adapting his strategy to achieve his objectives. This has several implications.

First, it is in practice seldom meaningful to try to capture the likelihood of intended harms in probabilistic terms. Instead of assigning probabilities to various acts by a terrorist, it is better to try to figure out what actions would best achieve the terrorist's objectives. In such an analysis, the terrorist's responses to one's own preparative defensive actions will have to be taken into account (Parnell et al. 2008). Game theory (that operates without probabilities) is better suited than traditional probability-based analyses to guide prevention aimed at reducing vulnerability to terrorist attacks and most other intentional threats (Hansson 2010).

Secondly, as noted by Golany et al. (2009), whereas the criterion of effectiveness is adequate in safety work, in security work it should be replaced by the criterion of vulnerability. Vulnerability can be understood as a weakness that can be exploited by an adversary. The adversary's aim is related to this loss and can in many cases be described as maximizing the loss (e.g., by targeting critical infrastructure). The optimal protection against terrorist attacks thus involves strategies to reduce the potential for loss. Probabilities do not have a central role in deliberations on how best to achieve such a reduction.

Sarewitz et al. (2003) add force to this line of argument by pointing out that vulnerability reduction can be considered a human rights issue, which may in some situations give it priority over economic optimization. Since modern society has an obligation to ensure that all citizens are provided a basic level of protection and that their fundamental rights are respected, economic arguments should not always be decisive in resource allocation. The authors give the example of the Americans with Disabilities Act (ADA), which requires that all public buses be provided with wheelchair access devices. This requirement was first opposed on economic grounds. Cost-benefit analyses showed that providing the buses with wheelchair access devices would be more expensive than providing, at public expense, taxi services for people with disabilities. The measure was nevertheless introduced, in order to realize the right of people with disabilities to be fully integrated into society. The right to protection against violence can be seen as a similar fundamental right, to be enjoyed by all persons. Such a right can justify protection even when a PRA or a CBA indicates that the resources would be "better" used elsewhere.

## Experiences and Examples

A discipline in which both approaches to the design for safety are being used is civil engineering and hydraulic engineering in particular. Civil engineering has a long history of applying safety engineering principles, in particular safety factors that have much of their origin in this domain of engineering (Randall 1976). Probabilistic risk analysis has a small but increasing role in civil engineering, most notably in the form of design criteria based on probabilistic information.

An example of how the safety factor approach is being used in hydraulic engineering is the geotechnical design of river dykes. One of the potential failure mechanisms of a slope is the occurrence of a slip circle, i.e., a rotational slide along a generally curved surface (Fig. 2).

The classic approach to determine the stability of a slope against sliding is to calculate for all possible sliding circles, the moment caused by the driving or destabilizing forces (i.e., the moment caused by the vertical arrows in Fig. 3) and the moment caused by the resisting or stabilizing forces (i.e., the moment caused by the leftward-directed arrows in Fig. 3). A slope is considered stable (or safe) if the ratio of the resisting momentum and the driving momentum is larger than a predefined safety factor. This safety factor is soil dependent. If the ratio is lower than the required safety factor, a flatter slope should be chosen and the calculation should be repeated. All engineers working with geotechnical materials are familiar with this iterative process of determining the maximum slope level (Terzaghi et al. 1996).

The landmark example of probabilistic design in hydraulic engineering is the design of the Dutch Eastern Scheldt storm surge barrier in the 1970s and 1980s. This was the last part of the Dutch Delta works, which were built in response to the severe North Sea flood of 1953. The original plan was to close off the Eastern

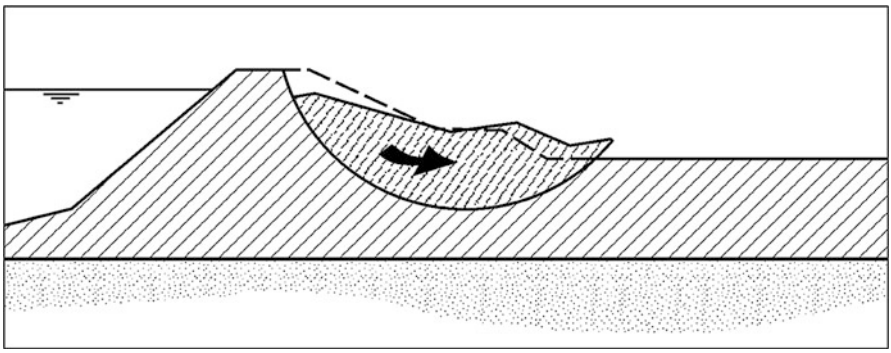


**Fig. 2** Slope instability of the inner slope (Source: TAW 2001; Kanning and Van Gelder 2008)

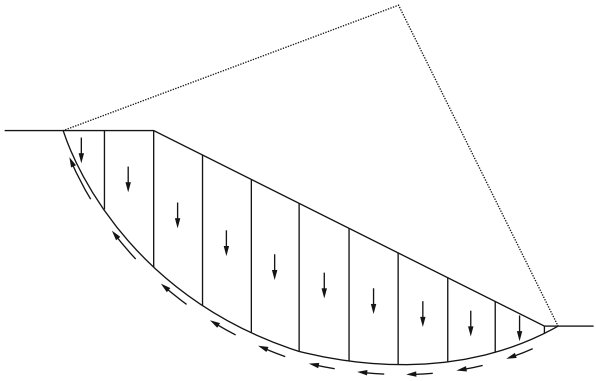


**Fig. 3** Potential sliding circle for geotechnical structure

Scheldt, but by the late 1960s, both environmentalists and fishermen opposed the full closure of the Eastern Scheldt. As an alternative, a storm surge barrier was designed that would normally be open and allow water to pass through, but that would close in case the water level at seaside exceeded a certain level.

The Eastern Scheldt storm surge barrier is the first hydraulic construction where the probabilistic design approach has been applied. Contrary to the design tradition at that time, the design process started with the construction of a fault tree, including as many failure mechanisms as possible, including failure of the system due to operation errors (CalIe et al. 1985).

According to Dutch water law, the Eastern Scheldt storm surge barrier had to be designed for 1/4,000-year conditions. This criterion specifies that the barrier has to be designed for a surge level and wave conditions that are expected to occur once every 4,000 years. Initially, this criterion was interpreted in terms of a 1/4,000-year design high water level at the seaside of the barrier, i.e., a water level that is expected to be exceeded only once every 4,000 years, together with 1/4,000-wave conditions. It was assumed that this design high water level at the seaside in combination with a low water level at landside of the barrier and extreme wave

conditions would determine the design load on the barrier. The result was a very unlikely combination of water level and wave conditions. It was therefore chosen to look at the combined 1/4,000-year hydraulic conditions, that is, the combination of high water level and wave conditions that had a probability of 1/4,000 years. This led to a reduction of 40 % in hydraulic load, as a result of which the distance between the pillars, an important design parameter, could be enlarged from 40 to 45 m. Similarly, some redundant elements in the design were removed because they did not significantly add to the overall safety (e.g., the removal of a backup valve for every opening in the barrier). However, the probabilistic approach did not only lead to the removal of redundant components of the barrier. On the basis of the fault tree, the weakest parts of the barrier were identified. Some elements were made stronger because that would significantly improve the overall safety (Vrijling 1990).

Despite recurrent pleas to switch from a "deterministic" to a probabilistic approach to design in hydraulic engineering, the prevalent design methodology is still based on the safety factor approach (Doorn and Hansson 2011). However, these safety factors are increasingly based on probabilistic calculations (Tsimopoulou et al. 2011). As such, they can be considered hybrid or mixed approaches. Probabilistic risk analysis approaches can play an important role after the design phase. Since probabilistic risk analysis approaches allow for comparison of the strengths of several elements within a system, they can accordingly indicate which element to improve. Therefore, probabilistic risk analysis approaches seem fit for identifying critical elements and setting up maintenance schemes (Vesely et al. 1994; Wang et al. 1996; Kong and Frangopol 2005). For the safety assessment of hydraulic structures after construction, probabilistic approaches increasingly replace the safety factor approach (Jongejan and Maaskant 2013; Schweckendiek et al. 2013).

## Critical Evaluation

In subsections "Arguments for Using Probabilistic Risk Analysis in Design" and "Arguments for the Safety Engineering Approach," we discussed the arguments in defense of design approaches using probabilistic risk analysis and design approaches that use principles from safety engineering, respectively. The strongest arguments in favor of design methods based on probabilistic risk analysis are the possibility of economic optimization and fitness for policy making (risk management). The strongest arguments for traditional safety engineering approaches refer to computational costs, simplicity, residual uncertainties, and security. Which approach is preferable when we want to design for safety? There is no general answer to that question; both approaches are of value, and it does not seem constructive to see them as competitors. In practice, neither of them can tell the full truth about risk and safety (Hansson 2009b). In order to see how we can combine the insights from both approaches, let us reconsider the objectives of the two approaches as explained in section "Current Approaches."

There are two different interpretations of the failure probabilities calculated in a PRA. One of these treats the calculated probabilities as relative indices of probabilities of failure that can be compared against a target value or against corresponding values for alternative designs. This interpretation seems unproblematic. It should be realized that it refers to a relative safety level; not all elements are included so it does not correspond to frequencies of failure in the real world (Aven 2009). Instead, this interpretation provides "a language in which we express our state of knowledge or state of certainty" (Kaplan 1993). It can be used to compare alternative components within a system, to set priorities or to evaluate the effects of safety measures. It is in such contexts of local optimization that probabilistic analysis has its greatest value (Lee et al. 1985).

The other interpretation treats the outcomes of PRA as objective values of the probability of failure. According to this view, these probabilities are more than relative indicators; they are (good estimates of) objective frequencies. In a world with no uncertainties but only knowable, quantifiable risks, this could indeed be a valid assumption. However, we do not live in such a world. In practice, failure probabilities of technological systems usually include experts' estimates that are unavoidably subjective (Caruso et al. 1999). Often some phenomena are excluded from the analysis. Such uncertainties make comparisons between different systems unreliable and sometimes severely misleading. To compare the safety of a nuclear power plant with the safety of a flood defense system on the basis of PRAs of the two systems is an uncertain and arguably not even meaningful exercise since the uncertainties in these two technologies are different and difficult or perhaps even impossible to compare.

Let us return to the safety factor approach in safety engineering that was said to be intended for compensating for five major categories of sources of failure (section "Safety Engineering"). Two of these, namely, higher loads and worse material properties than those foreseen, are targeted both by safety factors and probabilistic risk analysis. Due to the higher precision of probabilistic approaches, quantitative analysis of these sources of failure should at least in many cases preferably be based on probabilistic information.

The main advantage of the safety factor approach over probabilistic risk analysis concerns the other three sources of failure: imperfect theory of the failure mechanisms, possibly unknown failure mechanisms, and human error (e.g., in design). Probabilistic risk analysis is not capable of capturing these uncertainties. This is a major reason why probabilistic risk analysis should be seen as one of several tools for risk assessment and not as a sure source of final answers on risk assessment.

The more ignorant designers are of the uncertainties involved, the more they should rely on the traditional forms of safety engineering. Conversely, when uncertainty is reduced, the usefulness and reliability of probabilistic design methods is increased. There are currently no empirical standards regarding the appropriate design approach for different situations. It is desirable to carry out some action-guiding experiments to systematically evaluate the effect of the different approaches on the safety of a particular design.

## Conclusions

Probabilistic risk analysis is sometimes seen as competitor of traditional forms of safety engineering. This is a too narrow view of the matter. Neither of these methods can in practice tell the full truth about risk and safety. It is more constructive to see them as complementary. Probabilistic risk analysis is often an indispensable tool for priority setting and for the effect evaluation of safety measures. On the other hand, some of the uncertainties that safety engineering deals with successfully tend to be neglected in probabilistic calculations. Methodological pluralism, rather than monopoly for one single methodology, is to be recommended. Currently there is a trend in several fields of engineering towards increased use of probabilistic risk analysis. This trend will strengthen safety engineering, provided that it leads to a broadening of the knowledge base and not to the exclusion of the wide range of dangers – from one's own miscalculations to terrorist attacks – for which no meaningful probability estimates can be obtained.

## Cross-References


- ▶ Design for Values in Agricultural Biotechnology
- ▶ Design for Values in Engineering
- ▶ Design for Values in Nanotechnology
- ▶ Design for Values in Nuclear Technology
- ▶ Design Methods in Design for Values
- ▶ Modeling for Design for Values
- ▶ Design for Values and Operator Roles in Sociotechnical Systems


## References


Abbas AM, Manohar CS (2002) Investigations into critical earthquake load models within deterministic and probabilistic frameworks. Earthquake Eng Struct Dyn 31(4):813–832

Ale B (2009) Risk: an introduction. Routledge, London

Aven T (2009) Perspectives on risk in a decision-making context – review and discussion. Saf Sci 47(6):798–806

Blockley DI, Godfrey PS (2000) Doing it differently. Thomas Telford, London

CalIe EOF, Dillingh D, Meermans M, Vrouwenvelder AWCM, Vrijling JK, De Quelerij L, Wubs AJ (1985) Interim rapport TAW 10: Probabilistisch Ontwerpen van Waterkeringen. Technische Adviescommissie voor de Waterkeringen (TAW), Delft

Caruso MA, Cheok MC, Cunningham MA, Holahan GM, King TL, Parry GW, Ramey-Smith AM, Rubin MP, Thadani AC (1999) An approach for using risk assessment in risk-informed decisions on plant-specific changes to the licensing basis. Reliab Eng Syst Saf 63(3):231–242

Chapman PM, Fairbrother A, Brown D (1998) A critical evaluation of safety (uncertainty) factors for ecological risk assessment. Environ Toxicol Chem 17(1):99–108

Clausen J, Hansson SO (2007) Eurocodes and REACH: differences and similarities. Risk Manage 9(1):19–35

Clausen J, Hansson SO, Nilsson F (2006) Generalizing the safety factor approach. Reliab Eng Syst Saf 91(8):964–973

Council NR (1983) Risk assessment in the federal government: managing the process. National Academy Press, Washington, DC
Davis M (2001) Three myths about codes of engineering ethics. IEEE Technol Soc 20(Fall):8–14
Dhillon BS (1997) Failure mode and effects analysis: bibliography. Microelectr Reliab 32(5):719–731
Ditlevsen O, Madsen HO (2007[1996]) Structural reliability methods (internet edition 2.3.7). Wiley, Chichester
Doorn N, Hansson SO (2011) Should probabilistic design replace safety factors? Philos Technol 24(2):151–168
Fairbrother A (2002) Risk assessment: lessons learned. Environ Toxicol Chem 21(11):2261–2263
Golany B, Kaplan EH, Marmur A, Rothblum UG (2009) Nature plays with dice – terrorists do not: allocating resources to counter strategic versus probabilistic risks. Eur J Oper Res 192(1):198–208
Graham J, Wiener J (1995) Risk versus risk. Harvard University Press, Cambridge, MA
Guikema SD, Paté-Cornell ME (2002) Component choice for managing risk in engineered systems with generalized risk/cost functions. Reliab Eng Syst Saf 78(3):227–238
Hansson SO (1989) Dimensions of risk. Risk Anal 9(1):107–112
Hansson SO (1998) Setting the limit: occupational health standards and the limits of science. Oxford University Press, New York
Hansson SO (2009a) From the casino to the jungle. Synthese 168(3):423–432
Hansson SO (2009b) Risk and safety in technology. In: Meijers AWM (ed) Handbook of the philosophy of science. Philosophy of technology and engineering sciences, vol 9. Elsevier/North-Holland, Amsterdam, pp 1069–1102
Hansson, SO (2010) Promoting inherent safety. Process Safety and Environmental Protection Vol. 88(3), pp. 168–172
JCSS (2001) Probabilistic model code. Part 1 – BASIS of design. Joint Committee on Structural Safety. ISBN:978-3-909386-79-6
Jongejan RB, Maaskant B (2013) Applications of VNK2: a fully probabilistic risk analysis for all major levee systems in The Netherlands. In: Klijn F, Schweckendiek T (eds) Comprehensive flood risk management: research for policy and practice. Taylor & Francis, London, pp 693–700
Kanning W, Van Gelder PHAJM (2008) Partial safety factors to deal with uncertainties in slope stability of river dykes. In: De Rocquigny E, Devictor N, Tarantola S (eds) Uncertainty in industrial practice: a guide to quantitative uncertainty management. Wiley, London
Kaplan S (1993) Formalism for handling phenomenological uncertainties. The concepts of probability, frequency, variability, and probability of frequency. Nucl Technol 102(1):137–142
Keynes JM (1921) A treatise on probability. Macmillan, London
Knight FH (1935[1921]) Risk, uncertainty and profit. Houghton Mifflin, Boston
Knoll F (1976) Commentary on the basic philosophy and recent development of safety margins. Can J Civil Eng 3(3):409–416
Kong JS, Frangopol DM (2005) Probabilistic optimization of aging structures considering maintenance and failure costs. J Struct Eng-Asce 131(4):600–616
Lee WS, Grosh DL, Tillman FA, Lie CH (1985) Fault tree analysis, methods, and applications – a review. IEEE Trans Reliab 34(3):194–203
Mak KK, Sicking DL, Zimmerman K (1998) Roadside safety analysis program – a cost-effectiveness analysis procedure. Gen Des Roadside Saf Features 1647:67–74
Melchers RE (2002) Probabilistic risk assessment for structures. Proc Inst Civil Eng-Struct Build 152(4):351–359
Melchers RE (2007) Structural reliability theory in the context of structural safety. Civil Eng Environ Syst 24(1):55–69
Michal R (2000) The nuclear news interview. Apostolakis: on PRA. Nucl News 43(3):27–31
Miller CO (1988) System safety. In: Wiener EL, Nagel DC (eds) Human factors in aviation (cognition and perception). Academic, San Diego, pp 53–80

Möller N, Hansson SO, Peterson M (2006) Safety is more than the antonym of risk. J Appl Philos 23(4):419–432
Moses F (1997) Problems and prospects of reliability-based optimization. Eng Struct 19(4):293–301
Overton T, King GM (2006) Inherently safer technology: an evolutionary approach. Process Saf Progr 25(2):116–119
Parnell GS, Borio LL, Brown GG, Banks D, Wilson AG (2008) Scientists urge DHS to improve bioterrorism risk assessment. Biosecur Bioterror 6(4):353–356
Paté-Cornell ME (1996) Uncertainties in risk analysis: six levels of treatment. Reliab Eng Syst Saf 54(2–3):95–111
Rackwitz R (2004) Optimal and acceptable technical facilities involving risks. Risk Anal 24(3):675–695
Randall FA (1976) The safety factor of structures in history. Prof Saf 12–28
Rasmussen NC (1975) Reactor safety study. An assessment of accident risks in U.S. commercial nuclear power plants (WASH-1400, NUREG 75/014). U.S. Nuclear Regulatory Commission
Santillo D, Stringer RL, Johnston PA, Tickner J (1998) The precautionary principle: protecting against failures of scientific method and risk assessment. Mar Pollut Bull 36(12):939–950
Sarewitz D, Pielke R, Keykhah M (2003) Vulnerability and risk: some thoughts from a political and policy perspective. Risk Anal 23(4):805–810
Schweckendiek T, Calle EOF, Vrouwenvelder AWCM (2013) Updating levee reliability with performance observations. In: Klijn F, Schweckendiek T (eds) Comprehensive flood risk management: research for policy and practice. Taylor & Francis, London, pp 359–368
Takewaki I (2005) A comprehensive review of seismic critical excitation methods for robust design. Adv Struct Eng 8(4):349–363
TAW (2001) Technisch Rapport Waterkerende grondconstructies: Geotechnische aspecten van dijken, dammen en boezemkaden. Technische Adviescommissie voor de Waterkeringen (TAW)/Expertise Netwerk Water (ENW), Delft
Tench WH (1985) Safety is no accident. Collins/Sheridan House, London
Terzaghi K, Peck RB, Mesri G (1996) Soil mechanics in engineering practice, 3rd edn. Wiley, London
Tsimopoulou V, Kanning W, Verhagen HJ, Vrijling JK (2011) Rationalization of safety factors for breakwater design in hurricane-prone areas. Coastal structures 2011: Proceedings of the 6th international conference on coastal structures, Yokohama. World Scientific
Van de Poel IR, Royakkers LMM (2011) Ethics, technology, and engineering: an introduction. Wiley-Blackwell, West-Sussex
Vesely WE, Belhadj M, Rezos JT (1994) PRA importance measures for maintenance prioritization applications. Reliab Eng Syst Saf 43(3):307–318
Vrijling JK (1990) Kansen in de Waterbouw (inaugural address). Technical University Delft, Delft
Vrijling JK, van Hengel W, Houben RJ (1998) Acceptable risk as a basis for design. Reliab Eng Syst Saf 59(1):141–150
Vrouwenvelder A (2002) Developments towards full probabilistic design codes. Struct Saf 24(2–4):417–432
Wang J, Yang JB, Sen P, Ruxton T (1996) Safety based design and maintenance optimisation of large marine engineering systems. Appl Ocean Res 18(1):13–27
Zegeer CV, Carter DL, Hunter WW, Stewart JR, Huang H, Do A, Sandt L (2006) Index for assessing pedestrian safety at intersections. Transportation Research Record, No. 1982: Pedestrians and Bicycles. Transportation Research Board. National Academy of Sciences, Washington, DC, pp 76–83

# Design for the Value of Sustainability

Renee Wever and Joost Vogtländer

## Contents



**Abstract**

It is the main task of a professional designer to create value for the users of the products, services, and systems they design. In Design for Sustainability, however, designers have a higher level of ambition: additional to a high consumer value, they make sure that designs result in less degradation of our environment, less depletion of materials, and more social equity in our world. The need for a

R. Wever (✉)
TU Delft, Delft, The Netherlands

University of Limerick, Limerick, Ireland
e-mail: r.wever@tudelft.nl

J. Vogtländer
TU Delft, Delft, The Netherlands
e-mail: j.g.vogtlander@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_20

higher level of prosperity for people in developing countries, in combination with the growing population in our world, emphasizes the need for sustainable products and services. Design for Sustainability combines a high customer value with a low level of eco-burden over the life cycle. This chapter summarizes the main current approaches to Design for Sustainability (cradle-to-cradle, Circular Economy, and Biomimicry) and some practical tools and checklists (EcoDesign, the LiDS Wheel, Design for Recycling, and Design for Disassembly) and describes the latest developments in quantitative assessment methods ("Fast Track" Life Cycle Assessment, Eco-efficient Value Creation, and design of Sustainable Product Service Systems). For the quantitative methods, real-life examples are given for design of luxurious products based on cork, packaging design of food products, and Sustainable Product Service System design of sustainable water tourism.

**Keywords**

Life cycle assessment • Sustainability • Ecodesign • Eco-costs • Value • Product service systems

## Introduction

The current socioeconomic systems have brought us to an ever-increasing prosperity. This development will, however, inevitably come to an end because of its inability to stop the pollution of air, water, and soil and the degradation of ecosystems, to stop the depletion of material resources, and to support a growing world population in combination with the need for higher standards of living in the underdeveloped countries. The challenge of our generation is therefore to decouple societal progress from environmental deterioration and the use of nonrenewable resources. We need a better system of production and consumption to resolve this challenge.

The shaping of such a "new economy" requires high-level political decisions with respect to governmental regulations (on a national as well as a global scale). Companies, however, play an important role in the transition as well. They must serve their clients' needs with innovative high-value products and services, which cause less pollution. They must redesign their business systems in order to resolve the problem of materials depletion. This is not only an organizational challenge but also a challenge to designers and engineers, who must shape these new products and product service systems. The value for our society of Design for Sustainability is to support the required transition.

The purpose of this chapter is to provide an overview of the different approaches taken by designers for designing for the value of sustainability and of the complications and trade-offs they are likely to encounter. This overview will enable readers to better understand and place the work of specific designers, as well as the debates and critiques coming out of the design community. Readers will also have a better insight into the strengths and weaknesses of the assessment tools by which designers substantiate their decisions on sustainability.

## Sustainability and Design for Sustainability Explained

Sustainability is defined as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs" (Brundtland 1987). The issue is a matter of equity. It is about "intergenerational equity" (Tobin 1974), i.e., the notion that our children and grandchildren must have the same quality of our environment as we have (Gosseries 2008). It is also about "intragenerational equity," i.e., the equity within our own generation related to the poor countries of our world, the people of the so-called Base of the Pyramid (Prahalad 2002).

Although this definition of sustainability is widely accepted, its problem is that it defines sustainability in general terms on a global system level. Additional requirements and objectives are needed to translate the meaning of sustainability to goals and requirements for designing products and services.

A widely accepted approach toward capturing sustainability is by means of the Triple-P model (Elkington 1997). The term Triple-P is related to the aims of companies and to the design of products and services. According to this model (also called the Triple Bottom Line), equal weight in corporate activities should be given to the following three aspects:

- "People," the social aspects of employees in a company
- "Planet," the ecological consequences of the products of a company
- "Profit," the economic profitability

The main message is that the "bottom line" of an organization is not only an economic-financial one: an organization is responsible for its social and ecological environment as well. From this Triple-P perspective, an organization needs to find a balance between economic goals and goals with regard to the social and ecological environment.

At the World Summit on Sustainable Development 2002 in Johannesburg, "Profit" was changed in "Prosperity," and the emphasis of "People" shifted from the employees in a company to the "People of the Base of the Pyramid." So the model was brought in line with the report of Brundtland (1987), as depicted in Fig. 1. This figure also shows the relationship with Life Cycle Assessment (LCA) as described in section "Life Cycle Assessment ("Fast Track")" and the model of the Eco-costs/Value Ratio (EVR) as described in section "Eco-efficient Value Creation."

The original idea is that companies (and designers) must take well-balanced decisions on the 3 Ps. These decisions are considered as ones of making trade-offs between two sets of conflicting perspectives:

- Long term versus short term (Planet is long term, Profit is short term)
- "They" versus "us" in terms of the distribution of Prosperity (the People of the Base of the Pyramid versus consumers in developed countries)

Although the original idea of the Triple-P model was to make the right trade-offs in decision making, a more challenging way to approach the sustainability problem

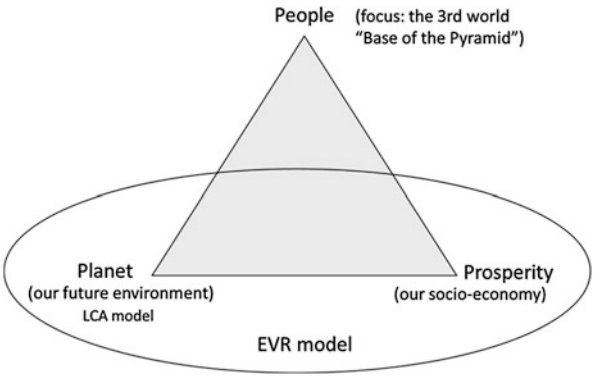



**Fig. 1** The Triple-P model for sustainability and its relationship with LCA and the EVR model (Source Vogtländer et al. (2013))

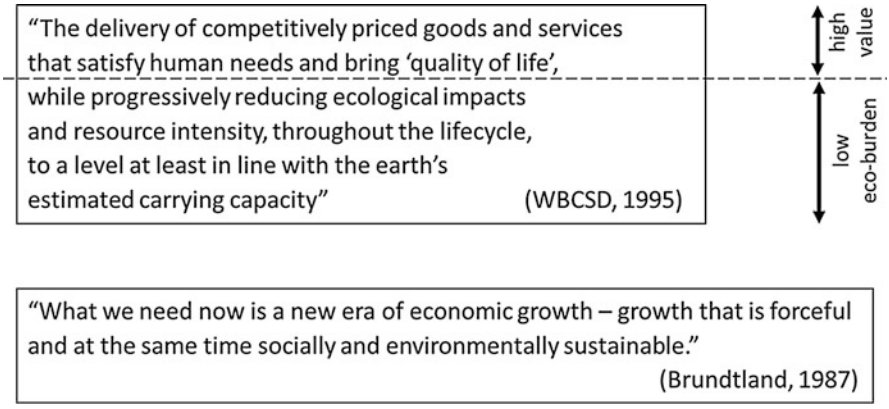



**Fig. 2** The corporate mission statement WBSCD (1995) and the conclusion of Brundtland (1987)

is not in terms of "or" but in terms of "and." This idea is called the "decoupling" or "delinking" of ecology and economy. This decoupling can be found in the general mission statement of the World Council for Sustainable Development, WBCSD, defined in November 1993 for their member companies, based on the double objective of the P of prosperity (created by product and services with a high added value) and the P of planet (low eco-burden), see Fig. 2.

The idea that such a double objective is the key to sustainable development was mentioned in the report of Brundtland (1987) as well: a new socioeconomic system is needed, see Fig. 2. Such a new economy can be made possible by innovation of products and services that combine a low eco-burden with a high value. This is the double objective of a modern designer.

But how do these general mission statements translate to the practical decisions designers and engineers have to make? There is a need to relate these statements to simple questions like the following: What is the best product, service, or system in terms of ecological impact? How can the designer improve the sustainability aspects of a design? What kind of solutions are available for what kind of situations?

This chapter on Design for Sustainability describes how sustainability is being translated to engineering and design practice. It deals with the following issues:

- Sustainable design approaches (section "Modern Holistic Sustainable Design Approaches")
- Tools and checklists to assist the designer in the quest for improvements (section "Tools and Checklists to Assist the Designer")
- Quantitative methods to assess the level of sustainability (section "Quantitative Methods to Assess the Level of Sustainability")

Cases are given in section "Examples."

The main current approaches to sustainable design are cradle-to-cradle and Circular Economy, described in sections "The Cradle-to-Cradle Approach" and "The Approach of the Circular Economy." The differences are limited: they both focus on elimination of the depletion of materials by recycling, obviously avoiding toxic emissions. They both claim a focus on "good rather than less bad" (claiming that they are better than previous approaches like eco-efficiency), but that appears to be a rather theoretical claim (starting from a envisioned ideal that may need to be watered down to make it achievable in the short run, instead of starting with the current reality and aiming to improve upon it), since what they achieve in practice often seems to be similar to other approaches. The main difference between cradle-to-cradle and Circular Economy is the emphasis the latter places on business model innovation.

These general approaches are providing a sustainable mind-set for designers, influencing the normal work of designers in a fundamental way. The designer should already apply these approaches at the beginning of the design (the so-called fuzzy front end) in situations where there is a high degree of design freedom.

Biomimicry, described in section "The Approach of Biomimicry," is a design method which aims at copying solutions from nature and is a logical "add-on" to cradle-to-cradle and Circular Economy. Obviously, inspiration from nature for technical solutions is far from new (it is as old as mankind) and is not restricted to the subject of sustainability, but it gives many designers inspiration on how to achieve a sustainable breakthrough. Furthermore, Biomimicry not only takes inspiration from nature for technical solutions but on a system level as well.

EcoDesign, also called Design for Environment (DfE) and Design for Sustainability (DfS), is briefly described in section "The Checklists of EcoDesign (Design for the Environment, Design for Sustainability)." Additional to the general approach of sustainable design, EcoDesign started with design manuals but is nowadays more and more web-based. Many computer software tools have been made (and are still being made) to assist the designer in decision making. It is applicable to all design stages and all kinds of degrees of design freedom. The LiDS Wheel, section "The LiDS Wheel (Environmental Benchmarking)," is part of EcoDesign but is presented separately since it is often used "stand-alone." EcoDesign checklists and tools are down-to-earth and very useful in practice.

One of the aspects covered in checklists deals with reduction of the impact during the use phase. This aspect, and then in particular the role of user behavior, is currently receiving a lot of research interest. These developments are described in section "Design for Sustainable Behavior" on design for sustainable behavior.

Design for Recycling and Design for Disassembly are related issues, to be dealt with in section "Design for Recycling and Design for Disassembly." This section provides practical design guidelines on what should be done to enable recycling. Design for Recycling has changed in the last decade to meet the requirements of modern waste separation techniques (shredding + materials separation).

The execution of all these approaches requires the ability to assess the impact of current products and systems as well as proposed alternatives. The most important quantitative method is Life Cycle Assessment (LCA), briefly described in section "Life Cycle Assessment ("Fast Track")." LCA is a product benchmarking tool, based on the material balance and the energy balance of a system. LCA of products is often called "cradle-to-grave," but in practice, it is in most of the cases cradle-to-cradle, since modern life cycles include recycling.

One of the drawbacks of LCA is that it only focuses on the Planet aspect of sustainability. An evolving quantitative design method is called Eco-efficient Value Creation. This method is LCA-based and is part of the model of the Eco-costs/Value Ratio. One of the key elements of Eco-efficient Value Creation (described in section "Eco-efficient Value Creation") is that innovative products must have a low eco-burden as well as a high customer value (both scoring better than the existing design solutions). Cases are given in section "Design of Luxurious Cork Products" (on luxurious design products of cork) and section "Packaging Solutions for Food" (on packaging). The Eco-efficient Value Creation design method is fully in line with the aforementioned approach of Brundtland (1987) and of WBCSD (1995).

Related to Eco-efficient Value Creation is the design of Sustainable Product Service Systems, SusPSS. A SusPSS is considered to be an optimum way to fulfill functional requirements with a minimum of eco-burden (in most cases with a minimal use of materials). In practice, the design of a SusPSS is often required to introduce green product solutions in the marketplace. Section "The Sustainable Product Service System (SusPSS)" describes when to design a SusPSS and how to do it. A case on sustainable water tourism is given in section "Sustainable Water Tourism, an Example of a SusPSS."

## Existing Approaches and Tools

### Modern Holistic Sustainable Design Approaches

#### The Cradle-to-Cradle Approach

Cradle-to-Cradle (McDonough and Braungart 2002) is a holistic view on how our socioeconomic system should be: away from the approach of minimization of the negative impact of our economic activities (eco-efficiency, making things that are

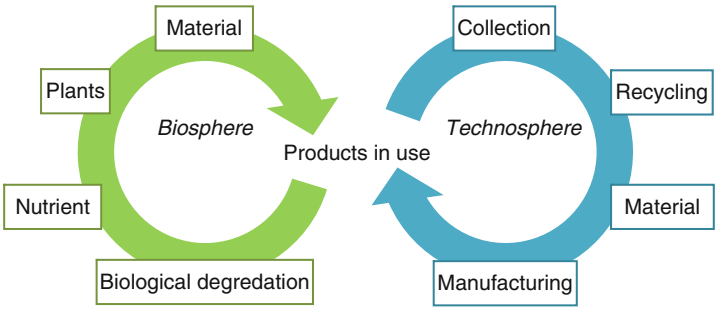



**Fig. 3** Material loops in the biosphere and in the technosphere according to cradle-to-cradle

less bad for our ecosystem) and toward stimulation and optimization of a positive impact (eco-effectiveness, making things that support our ecosystem). It describes a utopia with no restrictions to further growth of prosperity. Cradle-to-Cradle rejects the idea that economic growth is bad for our ecosystems: "...after all, in nature growth is good. Instead, it promotes the idea that good design supports a rich human experience with all that entails – fun, beauty, enjoyment, inspiration and poetry – and still encourages environmental health and abundance" (website MBDC 2013).

Cradle-to-Cradle takes the metabolisms in nature as an example: everything is recycled, with the sun as the energy source, where "waste is food," and where a high (bio)diversity exists. Materials in products can either be part of the biosphere (biodegradable materials) or be part of the technosphere where materials can continuously be upcycled (e.g., recycled to a level at least equal to the original quality), see Fig. 3.

The cradle-to-cradle principles are:

- Materials health, which means that materials which are used in products must be safe for use and continuous recycling. This means that they must be free of any potentially toxic substances.
- Materials reutilization, which means that all materials in the product must be part of continuous recycling systems.
- Renewable energy, which means that 100 % of the energy in the production and use phase must be renewable.
- Water stewardship, which means that water must be kept clean.
- Social fairness, which means that labor conditions must be kept at a high level.

### The Approach of the Circular Economy

The concept of the Circular Economy is based on cradle-to-cradle and several other previously existing approaches but is focused on the economy and on business opportunities. It is directly related with scientists who emphasized that the industrial economy should be reshaped from "linear" (based on fast replacement and disposal) to "circular" (reuse of products and recycling of materials) in order to stop depletion of resources (abiotic materials). Initial papers on the subject were

(Stahel 1982) and (Boulding 1966) both signaling that the current systems of production and consumption are not sustainable. Note that these papers were published before Brundtland (1987) and before cradle-to-cradle.

The Ellen MacArthur Foundation, established in 2010, revitalized the subject recently by combining the existing ideas of Industrial Ecology (clustering of industrial production to minimize waste), Design for Disassembly (section "Design for Recycling and Design for Disassembly"), Biomimicry (section "The Approach of Biomimicry"), and cradle-to-cradle (previous section). Moreover, the idea of the Circular Economy has been made appealing to the business community via reports on the economic importance and the business opportunities (McKinsey 2012, 2013). The timing is perfect: the European Union has put the issue of materials depletion high on the agenda (European Commission 2012).

The need to design products for the Circular Economy is simple: our current economy is not sustainable. Until now, the lifetime of products is getting shorter and shorter, and recycling rates of many materials are still far too low. The challenge can be explained by analyzing the following formula:

$$M = P \times W \times (1 - R) / L$$

where:

M = total required nonrenewable materials per year
P = total number of products = world population x average number of products per person
W = average weight of nonrenewable materials per product
R = fraction of reuse, remanufacturing, recycle, repair, refurbish, and retrieval (the so-called 6 Rs)
L = average lifetime of products in years

Our main problem is that P increases and that L decreases as a result of the growing world population, the increasing prosperity (in developing countries) and the trend of hyperconsumption. What can designers do?

1. The first and foremost task is to enhance L by products with a high level of quality and durability, which fulfill the user requirements, and are nice to have, so the owner will attach to them. Such a product has a high perceived customer value, so a high willingness to pay. This aspect is dealt with in section "Eco-efficient Value Creation."
2. The second task is to minimize W, by reducing weight, by applying renewables, and by optimizing the design in terms of fulfillment of the required functionality. This aspect is dealt with in sections "Life Cycle Assessment ("Fast Track")" and "The Sustainable Product Service System (SusPSS)."
3. The third task is enhancement of R, i.e., the 6 **R**s related to the issue of the Circular Economy, see Fig. 4:
   **R**euse (the product is sold on the secondhand market)

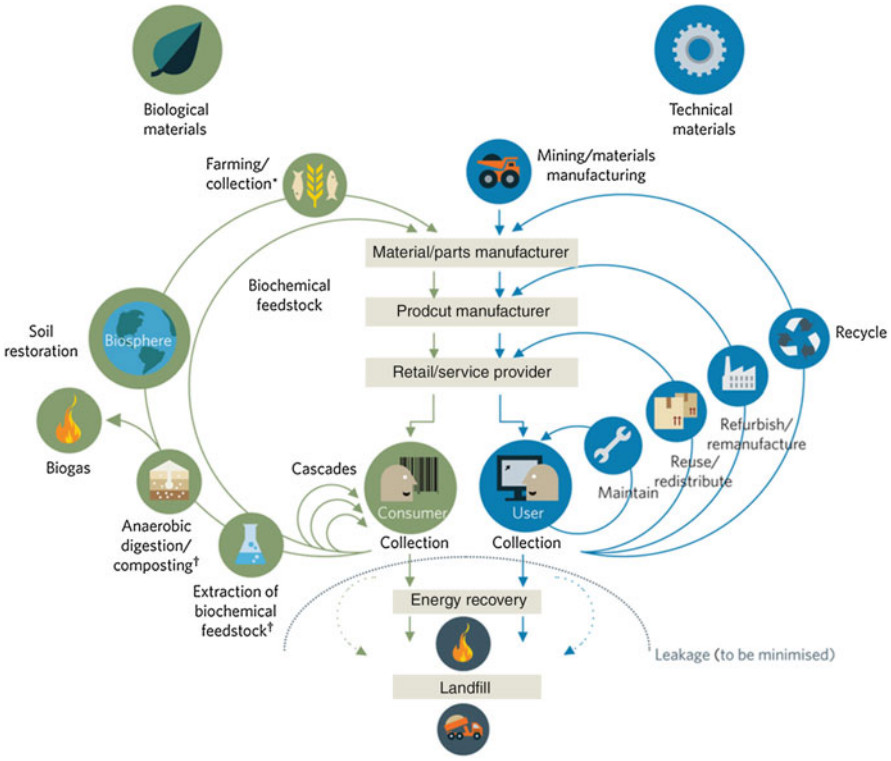



**Fig. 4** The Circular Economy in the biosphere and the technosphere (Copyright Ellen MacArthur Foundation)

**R**epair (the product life is extended by fixing its functionality)
**R**efurbish (the product life is extended by restoring its quality and image)
**R**emanufacturing (part of the product is remade)
**R**etrieval (part of the product is used in another product)
**R**ecycling (the materials are separated and either upcycled or downcycled in a cascade)

4. Make people want to live with less products (P), i.e., make a life with less "stuff" more desirable.

It is obvious that the designer can do a lot to make repair easy (related to Design for Disassembly, which will be dealt with in section "Design for Recycling and Design for Disassembly," keeping components with different life-spans separated (embedded batteries in smartphones and tablets are the wrong design trend!). The designer can also make the right choices with regard to postconsumer recycling, which is explained in section "Design for Recycling and Design for Disassembly."

There are three important issues in the discussions with regard to the introduction of Circular Economy product design:

- Products which are designed according to these principles appear often to be too expensive. This is a general issue with regard to green products and is dealt with in section "Eco-efficient Value Creation."
- New business structures are needed to introduce these products. This is also a general issue with regard to green products and is dealt with in section "The Sustainable Product Service System (SusPSS)."
- Customer behavior is an important aspect of optimizing their role in the Circular Economy. Here the challenge is to influence both the buying behavior of customers as well as influencing customer behavior at the end of the product life (e.g., stimulating waste separation by giving the customer easy access to a take-back system and/or giving the customer a small financial incentive for cooperation in the take-back system). Besides buying and discarding, there is the way people are using products. Section "Design for Sustainable Behavior" goes into more detail with regard to influencing the behavior during use.

## The Approach of Biomimicry

The understanding that nature has more effective life cycles than the technosphere makes Biomimicry a logical "add-on" to cradle-to-cradle and Circular Economy. Biomimicry is "innovation inspired by nature" (Benyus 1997). Copying from nature is as old as mankind, applied in modern mechanical and civil engineering (the honeycomb shape, the eggshell construction), architecture (the termite solution to keep buildings cool), and in aircraft design (miniaturization of airplanes). Simply copying a form and/or technical solution from nature is often referred to as biomimetics. In Biomimicry, the aim is to also copy nature at the process and system level, for instance with closed material loops. These higher levels are represented by working with the so-called Life's Principles (see below). The approach of Biomimicry has been embraced recently by designers who are inspired by the approaches of Cradle-to-Cradle and the Circular Economy and are looking for sustainable solutions in design and engineering. Biomimicry is a way to search for new sustainable materials, products, and systems. Examples are shown on the website Biomimicry (2013).

Biomimicry is more a holistic approach than a practical design tool. The inspiration must come from the design team, who can use the so-called Life's Principles as a guide. Life's Principles are design lessons from nature: life has evolved on Earth over several billion years. We might learn from these patterns. The Life's Principles are the summary of patterns which evolved to achieve optimal ecosystems and give practical guidance to the designer. The checklist:

- Adapt to changing conditions (incorporate diversity, maintain integrity through self-renewal, embody resilience through variation, redundancy, and decentralization).

- Be locally attuned and responsive (leverage cyclic processes, use readily available materials and energy, use feedback loops, cultivate cooperative relationships).
- Use life-friendly chemistry (break down products into benign constituents, build selectively with a small subset of events, do chemistry in water).
- Be resource efficient (material and energy) (use low-energy processes, use multifunctional design, recycle all materials, fit form to function).
- Integrate development with growth (self-organize, build from the bottom up, combine modular and nested components).
- Evolve to survive (replicate strategies that work, integrate the unsuspected, reshuffle information).

## Tools and Checklists to Assist the Designer

### The Checklists of EcoDesign (Design for the Environment, Design for Sustainability)

EcoDesign is meant to assist the designer in creating sustainable products, services, and systems, in all stages of the design process. It is making designers aware of what can be done to improve the design. UNEP (United Nations Environment Programme) published the first groundbreaking EcoDesign manual (Brezet and Van Hemel 1997). By experiences gained from its application in practice, EcoDesign evolved through Design for Environment (DfE) to the broader concept of Design for Sustainability (DfS), which includes the social issues of sustainability and the need to develop new ways to meet the consumer needs in a less resource-intensive way (Crul and Diehl 2006; Crul et al. 2009). These manuals are full of step-by-step procedures, design tools, checklists, and examples. Designers highly appreciate tools in the form of checklists, since that seems to help them during the design process. It makes them feel that nothing has been overlooked. It is beyond the purpose of this chapter to describe all these tools, but Fig. 5 gives an overview of such checklists in (Brezet and Van Hemel 1997).

The EcoDesign checklists of Fig. 5 consist of two columns: the questions to be asked are given in the left-hand columns of the tables. Some improvement options are suggested in the right-hand columns. The checklists are related to the LiDS Wheel which will be presented in the next section.

In practice, however, the application of EcoDesign manuals is rather limited. They provide a good basis for training on the subject, but evidence shows that designers rather "file" them than use them. Designers ask for easy accessible and inspiring examples, and easy accessible specific information, since they claim that they have limited time available: in design many other quality aspects come before "eco" (Lofthouse 2006). The obvious thing to do is to develop open-access databases on the Internet. Until now, information on the Internet is rather scattered, so it takes a lot of time to gather the inspiration and data which are needed. The situation is improving, however, step-by-step. There is still a long way to go.

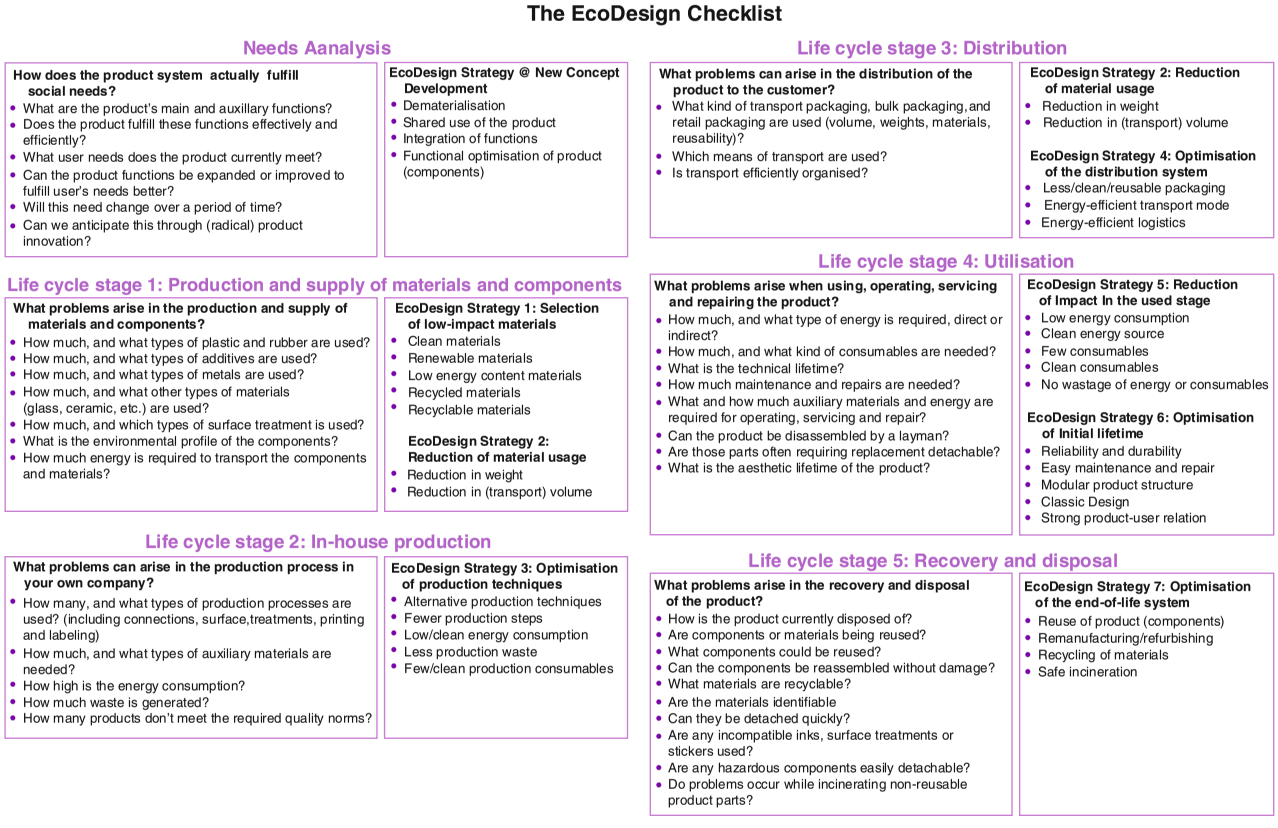



**Fig. 5** The EcoDesign checklist (As reproduced in Van Boeijen and Daalhuizen 2013)

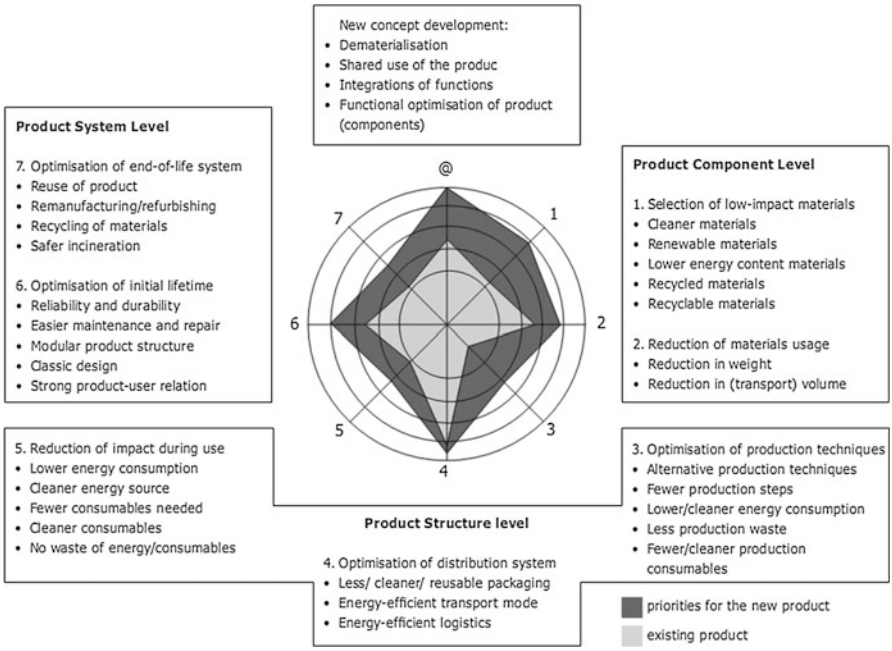



**Fig. 6** The LiDS Wheel (Brezet and Van Hemel 1997; Van Hemel 1998)

Another aspect of the limited application of EcoDesign is that the vast majority of designers claim that:

- The design brief does not ask for it or that many other aspects come first (e.g., Petala et al. 2010).
- The design becomes too expensive by EcoDesign (note that the claim that EcoDesign is less expensive is not supported by most of the practical cases: EcoDesign is definitely not a cost-saving tool).

It is evident that the design of sustainable products and services must be part of the marketing and production strategy of a company: it can only thrive in a company culture in which sustainability is fully embedded. The leadership of the company management plays a crucial role in successful implementation.

### The LiDS Wheel (Environmental Benchmarking)

The LiDS Wheel (Brezet and Van Hemel 1997; Van Hemel 1998) is a specific EcoDesign tool, which is part of nearly all EcoDesign methods, and which is widely used by consultants. It is also called EcoDesign Web (Bhamra and Lofthouse 2003), D4S Strategy Wheel (Crul and Diehl 2006; Crul et al. 2009), and EcoDesign Strategy Wheel (Van Boeijen and Daalhuizen 2013). LiDS is an abbreviation of Life Cycle Design Strategy. See Fig. 6.

Basically, it is a form of environmental product benchmarking, showing in what aspects a product design should be improved, compared to its alternatives. It only works in situations where two or more products (designs) are compared to each other, since scores are not absolute, but relative and often subjective. The advantage of the tool is that it is quick and dirty and the disadvantage is that the importance of each aspect relative to the other aspects is not known, which can easily lead to too much focus on the wrong aspects.

### Design for Sustainable Behavior

In the clarifying text of the fifth strategy of the LiDS Wheel "reduction of impact during use," the mentioned strategies all aim at technical optimization. However, a large part of the impact that products have during their life-span, especially energy- and water-using products, is a result of the behavior by their users. Think of the excess amount of water heated in an electric kettle or the energy consumed by products left in standby mode. The possibilities to influence such behavior for the better have received substantial attention from sociologists and psychologists over the years. However, those researchers have not seen design as a true variable (Wever 2012). In recent years, a vibrant area of design research has emerged studying the potential of influencing the user through the design to change their behavior in a more sustainable way (e.g., Lockton et al. 2008; Lilley 2009). As Bakker (2012) states, there are three design approaches to reduce the impact of (inefficient) use. The first is an engineering approach, aimed at automating certain aspects of products in order to increase efficiency, i.e., engineering away inefficiencies. The second is an individualist approach, which comes down to Design for Sustainable behavior, focusing on isolated, specific user-product interactions (e.g., dual-flush toilet buttons). The third is a practice approach, taking practices (comprised of material artifacts or "stuff," conventions or "image" and competencies or "skills") as the unit of study (e.g., the complete practice of bathing or cooking). Several scholars are working toward practical tools enabling designers to design for sustainable behavior change (Lockton et al. 2010; Daae and Boks 2013; Lilley and Wilson 2013).

### Design for Recycling and Design for Disassembly

Design for Recycling and Design for Disassembly are basically EcoDesign issues, but until now, they are a bit neglected by designers. It is about details in the design and about technical processes at the end-of-life, both not very popular design aspects. It is expected that this situation will change soon since, on the one hand, the issue of materials depletion is rapidly getting high on the agenda of all stakeholders, resulting in attention to the 6 Rs of the Circular Economy (see section "The Approach of the Circular Economy") and since, on the other hand, manufacturers will be confronted with their own poor product design with respect to recycling due to the introduction of obligatory product take-back systems in the European Union.

Design for Recycling is defined at the Life Cycle Thinking website of the Joint Research Centre of the European Commission (website JRC 2013) as: "... a method that implies the following requirements of a product: easy to dismantle, easy to obtain 'clean' material-fractions, that can be recycled (e.g., iron and copper should be easy to separate), easy to remove parts/components, that must be treated separately, use as few different materials as possible, mark the materials/polymers in order to sort them correct, avoid surface treatment in order to keep the materials 'clean'."

Design manuals in this field are rather specific, since only detailed descriptions can guide the designer what to do. Examples are Recoup (2009) for plastic bottles and Chiodo et al. (2011) for the active disassembly of products. Still, the problem for designers is that Design for Recycling is strongly related to complex issues about production and assembling (Boothroyd et al. 2002). Designers have to resolve problems in teams together with engineering and manufacturing, which is a challenge to many designers.

A common argument to refrain from the complexities of Design for Recycling is that it makes a product more expensive. As such, this is true, but the introduction of Design for Recycling is driven by the fact that industry must soon comply with strict governmental regulations within the European Union. As an example, detailed lists of substances which have to be removed prior to disposal can be found in the WEEE (Waste Electrical and Electronic Equipment) Directive of the European Union.

## Quantitative Methods to Assess the Level of Sustainability

### Life Cycle Assessment ("Fast Track")

A basic question in Design for Sustainability is which design (of a series of designs or concepts) is the best in terms of sustainability.[1] This is a matter of benchmarking and typically results in making trade-offs. When product A requires less electricity than product B, but product A requires more transport than B, such a trade-off arises. The issue of comparison of environmental burden of different aspects has already been mentioned in section "The LiDS Wheel (Environmental Benchmarking)" on the LiDS Wheel. It becomes even more complex when product A requires more of material X and less of material Y. The relative eco-burden per kilogram of X and Y is then required for a benchmarking analysis. This is where Life Cycle Assessment (LCA) comes in.

LCA is a well-defined method to calculate the environmental burden of a product or service. The basic calculation structure of LCA is depicted in Fig. 7. The calculation is based on a system approach of the chain of production and

[1]Parts of this section, including figures, have been copied from Chaps. 2.1, 2.6, and 3.1 of the LCA guide for students, designers, and business managers (Vogtländer 2012).

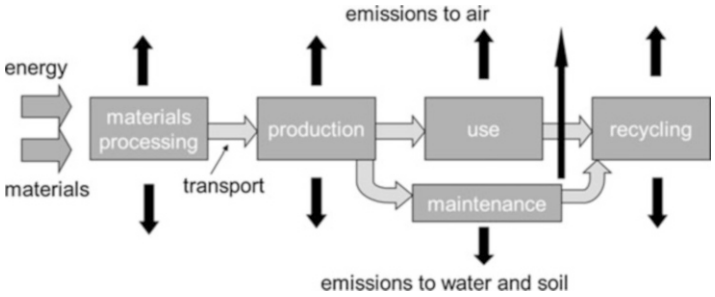



**Fig. 7** The basic calculation system of LCA

consumption, analyzing the input and the output of the total system by a materials balance and an energy balance of the total system:

- Input:
  - Materials (natural resources and recycled materials)
  - Energy
  - Transport
- Output:
  - The product(s) and/or service
  - Emissions to air, water, and soil
  - By-products, recycling products, feedstock for electrical power plants
  - Waste for landfill, waste incineration, or other types of waste treatment

Each LCA starts with the definition of the different processes inside the black boxes of Fig. 7. Such definitions are unique for each case. When the definition of the system, i.e., the combination of processes to be studied, is wrong (or not suitable for the goal of the study), the output of the calculation will be wrong as well. The biggest mistakes in practice are caused by a system definition which is too narrow: subprocesses are not included which may nevertheless be important, or other details are included which have hardly any influence on the output. The definition of the system is often iterative: by trial and error, it is discovered what is important in a certain case.

Some cradle-to-cradle specialists claim that the cradle-to-grave dogma of LCA leads to a wrong approach in design. They have a point that the cradle-to-grave dogma may lead to wrong design decisions (i.e., opportunities for recycling are overlooked). However, this has nothing to do with LCA methodology as such, but only with the people who apply it. Reuse, recycling, etc. are always part of the LCA of the whole system. Yet it is important to give ample attention in LCA to the definition of the system, its boundaries, and its function (Vogtländer 2012).

The application of LCA is totally different for two groups of users:

- The classical LCA ("full," "rigorous"), where the methodological focus is on the LCI (Life Cycle Inventory, i.e., making lists of emissions and required natural

resources) and on the LCIA (Life Cycle Impact Assessment, analyzing these lists and "compressing" such lists in "single indicators"), which is work for LCA specialists and scientists.
- The "Fast Track" LCA[2] is where the output of the calculations of the classical LCA is the input for the Fast Track calculation and where the methodological focus is not at all on the LCI and the LCIA but on the comparison of design alternatives, which is work for designers, engineers, and business managers.

The Fast Track LCA is designed to make LCA doable by the designers themselves. When a product is designed (e.g., a car, a house), all kinds of materials and production processes are combined. It is inconceivable that all these materials and processes are analyzed by the designer himself on the level of individual emissions and use of natural resources. In practice, the designer will apply the results of LCAs from other people, available in databases (e.g., the Idemat 2012 database or the Ecoinvent database with over 5,000 LCIs of different processes). There are also generic engineering tools which include basic sustainability assessment options such as technical documentation tools like SolidWorks or material selection tools such as the Cambridge Engineering Selector (CES).

Since the aim of a Fast Track LCA is a comparison of products, the first thing to do for carrying it out is to select a single indicator model. The most common models for single indicators are: ReCiPe (a so-called damage -based indicator, aiming at quantifying the ecological damage caused by processes, with points as the indicator unit), carbon footprint (a so-called single-issue indicator, since it deals only with greenhouse gasses, with kg $CO_2$ equivalent as the unit), cumulative energy demand (CED, embedded energy, megajoules as the unit), and eco-costs (a so-called prevention-based indicator, aiming at quantifying the cost of repairing caused damages, with euros as the unit). Eco-costs are "external costs," related to the marginal costs of prevention methods that have to be taken to bring the total emissions back to the "no-effect level" (restore the equilibriums in our ecosystems), so eco-costs are "hidden obligations." Eco-costs are the proxies of the tradable emission right levels that are required to resolve the problem of environmental degradation. As it is expected that governmental regulation will become ever stricter until the problem is solved, eco-costs can soon become "internal costs" and are risks for companies of future noncompliance with those regulations. Therefore, eco-costs have a direct relationship to the future profit of companies, so they are relevant for designers and business managers. See Fig. 8.

It is a widespread misunderstanding that a Fast Track LCA is less accurate than a rigorous classical LCA. The accuracy is not less, since it is based on formal databases and since it is calculated according to the general rules of LCA as described in handbooks (Guinée 2002; ILCD 2010; BSI 2011) and specified in

[2]Also called the "Philips method," since Philips Electronics was the first company which did LCAs in this way in 1998–1999 and developed the EcoScan software.

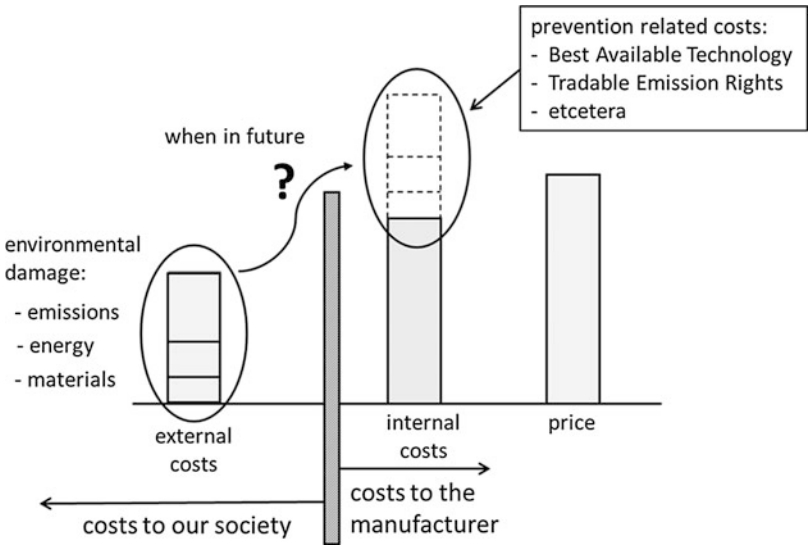



**Fig. 8** Eco-costs will gradually become internal costs as a consequence of governmental regulations; the question is not if but when

ISO 14040 (2006), 14044 (2006), and 14067 (2013). Fast Track LCA is made assessable to designers (Vogtländer 2012). Lookup tables are provided, as shown in Fig. 9.

The Fast Track LCA method has the following step-by-step procedure:

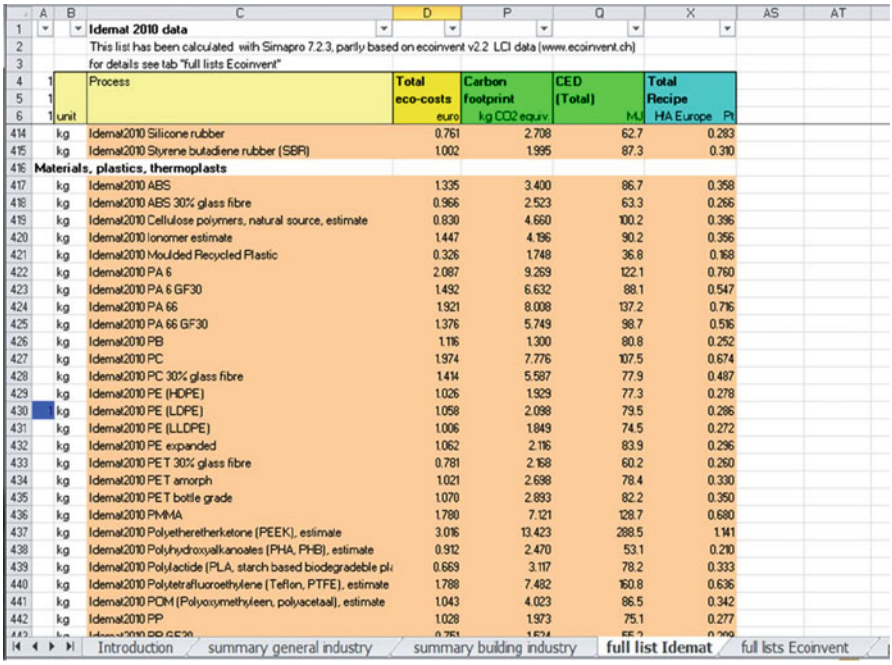
- Step 1: Establish the scope and the goal of your analysis (this step might be done after step 2 in the case that it is a total new design).
- Step 2: Establish the system, functional unit, and system boundaries.
- Step 3: Quantify materials, use of energy, etc. in your system.
- Step 4: Enter the data into an Excel calculation sheet or a computer program.
- Step 5: Interpret the results and draw your conclusions.

## Eco-efficient Value Creation

### The Model of the Eco-Costs/Value Ratio (EVR)

A prerequisite for a comparison in LCA (LCA benchmarking) is that the functionality ("functional unit") and the quality of the alternative product(s) are the same: you cannot compare apples and oranges with LCA.[3] In cases of product design and architecture, however, this prerequisite seems to be a fundamental flaw in the application of LCA: the designer or architect is aiming at a better functionality

[3]Parts of this section have been copied from Chap. 2 of the book on Eco-efficient Value Creation (Vogtländer et al. 2013).



| | | | C | D | P | Q | X | AS | AT |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | **Idemat 2010 data** | | | | | | |
| 2 | | | This list has been calculated with Simapro 7.2.3, partly based on ecoinvent v2.2 LCI data (www.ecoinvent.ch) | | | | | | |
| 3 | | | for details see tab "full lists Ecoinvent" | | | | | | |
| 4 | 1 | | Process | **Total eco-costs** | **Carbon footprint** | **CED (Total)** | **Total Recipe** | | |
| 5 | 1 | | | | | | | | |
| 6 | 1 | unit | | euro | kg CO2 equiv | MJ | HA Europe Pt | | |
| 414 | | kg | Idemat2010 Silicone rubber | 0.761 | 2.708 | 62.7 | 0.283 | | |
| 415 | | kg | Idemat2010 Styrene butadiene rubber (SBR) | 1.002 | 1.995 | 87.3 | 0.310 | | |
| 416 | **Materials, plastics, thermoplasts** | | | | | | | | |
| 417 | | kg | Idemat2010 ABS | 1.335 | 3.400 | 86.7 | 0.358 | | |
| 418 | | kg | Idemat2010 ABS 30% glass fibre | 0.966 | 2.523 | 63.3 | 0.266 | | |
| 419 | | kg | Idemat2010 Cellulose polymers, natural source, estimate | 0.830 | 4.660 | 100.2 | 0.396 | | |
| 420 | | kg | Idemat2010 Ionomer estimate | 1.447 | 4.196 | 90.2 | 0.356 | | |
| 421 | | kg | Idemat2010 Moulded Recycled Plastic | 0.326 | 1.748 | 36.8 | 0.168 | | |
| 422 | | kg | Idemat2010 PA 6 | 2.087 | 9.269 | 122.1 | 0.760 | | |
| 423 | | kg | Idemat2010 PA 6 GF30 | 1.492 | 6.632 | 88.1 | 0.547 | | |
| 424 | | kg | Idemat2010 PA 66 | 1.921 | 8.008 | 137.2 | 0.716 | | |
| 425 | | kg | Idemat2010 PA 66 GF30 | 1.376 | 5.749 | 98.7 | 0.516 | | |
| 426 | | kg | Idemat2010 PB | 1.116 | 1.300 | 80.8 | 0.252 | | |
| 427 | | kg | Idemat2010 PC | 1.974 | 7.776 | 107.5 | 0.674 | | |
| 428 | | kg | Idemat2010 PC 30% glass fibre | 1.414 | 5.587 | 77.9 | 0.487 | | |
| 429 | | kg | Idemat2010 PE (HDPE) | 1.026 | 1.929 | 77.3 | 0.278 | | |
| 430 | | kg | Idemat2010 PE (LDPE) | 1.058 | 2.098 | 79.5 | 0.286 | | |
| 431 | | kg | Idemat2010 PE (LLDPE) | 1.006 | 1.849 | 74.5 | 0.272 | | |
| 432 | | kg | Idemat2010 PE expanded | 1.062 | 2.116 | 83.9 | 0.296 | | |
| 433 | | kg | Idemat2010 PET 30% glass fibre | 0.781 | 2.168 | 60.2 | 0.260 | | |
| 434 | | kg | Idemat2010 PET amorph | 1.021 | 2.698 | 78.4 | 0.330 | | |
| 435 | | kg | Idemat2010 PET bottle grade | 1.070 | 2.893 | 82.2 | 0.350 | | |
| 436 | | kg | Idemat2010 PMMA | 1.780 | 7.121 | 128.7 | 0.680 | | |
| 437 | | kg | Idemat2010 Polyetheretherketone (PEEK), estimate | 3.016 | 13.423 | 288.5 | 1.141 | | |
| 438 | | kg | Idemat2010 Polyhydroxyalkanoates (PHA, PHB), estimate | 0.912 | 2.470 | 53.1 | 0.210 | | |
| 439 | | kg | Idemat2010 Polylactide (PLA, starch based biodegradeble pl | 0.669 | 3.117 | 78.2 | 0.333 | | |
| 440 | | kg | Idemat2010 Polytetrafluoroethylene (Teflon, PTFE), estimate | 1.788 | 7.482 | 160.8 | 0.636 | | |
| 441 | | kg | Idemat2010 POM (Polyoxymethyleen, polyacetaal), estimate | 1.043 | 4.023 | 86.5 | 0.342 | | |
| 442 | | kg | Idemat2010 PP | 1.028 | 1.973 | 75.1 | 0.277 | | |

Introduction | summary general industry | summary building industry | **full list Idemat** | full lists Ecoinvent

**Fig. 9** Screenshot of part of the lookup table for products, services, and energy (website ecocostsvalue 2013)

and quality (in the broad sense of the word: including intangible aspects like beauty and image), so the new design never has the same functionality and quality as the old solution. As an example, we look at an armchair: different types of armchairs differ in terms of comfort, aesthetics, etc. rather than in terms of the functionality of "providing support to sit." One solution is to take the market value (Willingness to Pay, WTP) as a proxy for the sum of all quality aspects (tangible as well as intangible) also into account and determine the ratio of the eco-burden (determined by LCA) and the value (in euros, US dollars, or any other currency). This leads to the Eco-costs/Value Ratio (EVR) as an indicator for the sustainability of a product.

The future trend of "internalizing" eco-costs, as depicted in Fig. 8, might be a threat to a company, but it might also be an opportunity: "When my product has less eco-burden than that of my competitor, my product can withstand stricter regulations of the government." So the characteristic of low eco-costs of products is a competitive edge in the future. To analyze the short-term and the long-term market prospects of a product or a product service combination (Product Service System, PSS), each product can be positioned in the product portfolio matrix of Fig. 10. *On the vertical axis are the eco-costs, on the horizontal axis is the ratio between value over production costs*, i.e., *an indicator for the current added value of the business activity. So here "costs" are the real costs, and "eco-costs" are the virtual costs representing the eco-burden of a product or service, with the understanding that*

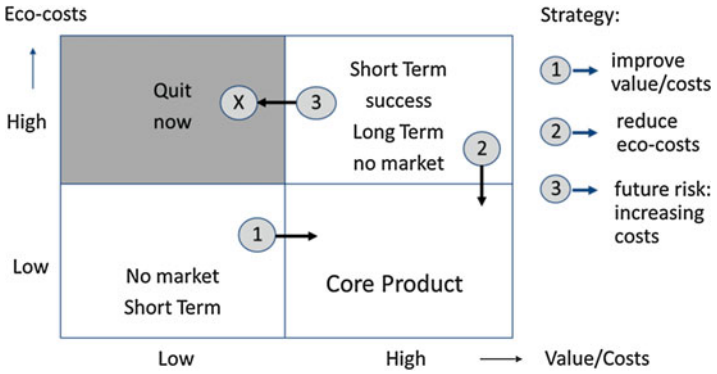



**Fig. 10** Sustainable Business Strategy Matrix for products of companies

*eco-costs may well become real costs in the future through legislation internalizing this burden (*i.e., *the polluter pays principle).*

The high value/costs Ratio gives the opportunity for a company of making high profits (by a high profit margin per product and/or by a high market share). Low eco-costs make that the good position in the market will not deteriorate in the future by stricter environmental regulations and higher material prices. So the lower right quadrant in the matrix is the desired position.

For many "green designs," the usual problem is that they have a low current value/costs ratio (the lower left quadrant). In most of the cases, the production costs of these green designs are higher than the production costs of the classic solution; in some cases, even the (perceived) quality is poor, so the value is lower than the classic solution. There are two ways to do something about it (arrow 1 in Fig. 10):

- Enhance the (perceived) quality of the product.
- Attach a service to the product (create a PSS) in a way that the value of the bundle of the product and the service is more than the value of its components (sections "The Sustainable Product Service System (SusPSS)" and "Sustainable Water Tourism, an Example of a SusPSS").

For most of the current products, the value/costs ratio is high, but the eco-costs are high as well (the upper right quadrant). Doing nothing is no option: it will cause products to drift into the upper left quadrant (arrow 3), because of the aforementioned "internalization" of costs of pollution. Such a product will be forced out of the market, since the sale price of the product cannot be increased above the fair price in the eyes of the customer (the Willingness to Pay). These products and its production processes have to be redesigned to lower the eco-costs (arrow 2).

There is also a consumer's side of the EVR model: the decoupling of economy and ecology (as mentioned in section "Introduction"). Under the assumption that most of the households spend in their life what they earn in their life (the savings ratio is <5 % in most countries), the total EVR of the spending of households is the

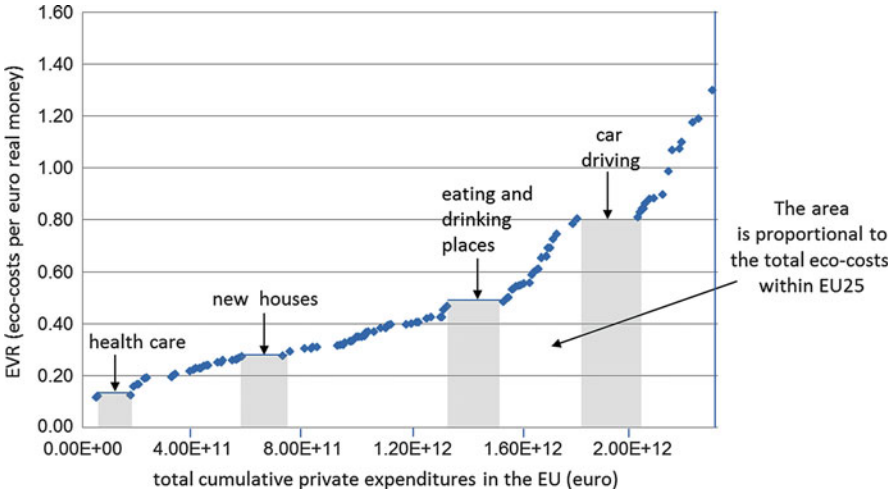



**Fig. 11** The EVR and the total expenditures of all consumers in 25 European Union countries (EU25)

key toward sustainability. Only when this total EVR of the spending gets lower, the eco-costs related to the total spending can be reduced even at a higher level of spending.

A short macroeconomic analysis on what happens in the European Union reveals what can be done. Figure 11 shows the EVR (=eco-costs/price) on the Y-axis as a function of the cumulative expenditures of all products and services of all citizens in the 25 countries that made up the European Union at the time on the X-axis, derived from the EIPRO study of the Joint Research Centre of the European Commission (Tukker et al. 2008).

The area underneath the curve is proportional to the total eco-costs of the 25 European Union countries. Basically there are two strategies to reduce the area under the curve:

- Force industry to reduce the eco-costs of their products (this will shift the curve downward), e.g., by cleaner and more energy efficient production, less transport, less energy in the use phase, and closed loop recycling.
- Try to reduce expenditures of consumers in the high end of the curve, and let them spend this money at the low end of the curve (this will shift the middle part of the curve to the right), e.g., tempt consumers to spend their money on health care and new houses, rather than on car driving.

Designers and engineers cannot only contribute to the first option but also to the second, by designing innovative products with a low EVR, which are attractive to the consumers (so that they will buy these products). These products must have a higher value (higher WTP) than the existing alternative they must replace, to avoid the so called “rebound effect” (Sorrell 2007).

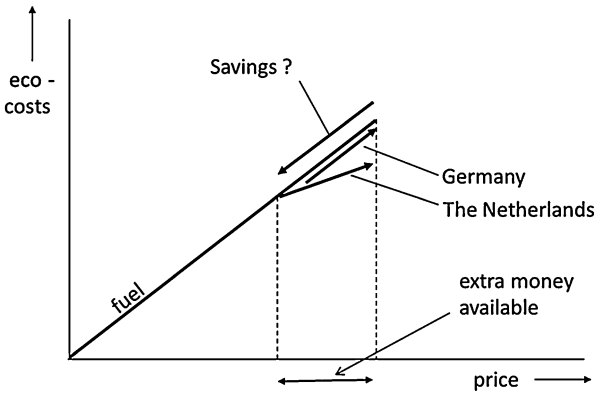



**Fig. 12** Reduction of the fuel consumption of a car by better aerodynamics, an example of the rebound effect

### The Rebound Effect

The rebound effect refers to increased consumption which results from actions that increase efficiency and reduce consumer costs (i.e., savings lead to expenditures, since we spend what we earn in our life). Three types of rebound effect can be distinguished:

- The *direct* rebound effect ("substitution effect") where the rebound is in the same function (e.g., people who install low-energy light bulbs tend to be less strict on turning off the light when they leave the room or even install these light bulbs in their gardens)
- The *indirect* rebound effect ("income effect") where the rebound is in another functional area (e.g., people tend to travel more when they save money by energy conservation)
- The *economic-wide, long-term* rebound effect (e.g., when cars become more energy efficient, more people can afford driving, resulting in more cars)

The rebound effect (direct and indirect) is an important issue in the model of the EVR. An example of a direct rebound effect from the automotive industry is given in Fig. 12.

On the first sight, Fig. 12 shows that better aerodynamics is a win-win situation in the use phase, since it results in savings in eco-costs (at a high EVR) as well as a lower price for the consumed fuel. However, the money saved on fuel is spent again on fuel in a country with no speed limit, like Germany. This is because of consumer preferences: the advantage of a better aerodynamics is transferred to driving faster (perceived as convenience and fun), instead of savings on diesel consumption (savings of eco-costs). In the Netherlands, a country with speed limits, the situation is slightly better. It results in driving more. "Driving more" has a lower EVR than "driving faster" (the EVR of the diesel is higher than the EVR for the car + diesel), but the end result is that there are not much overall savings in eco-costs.

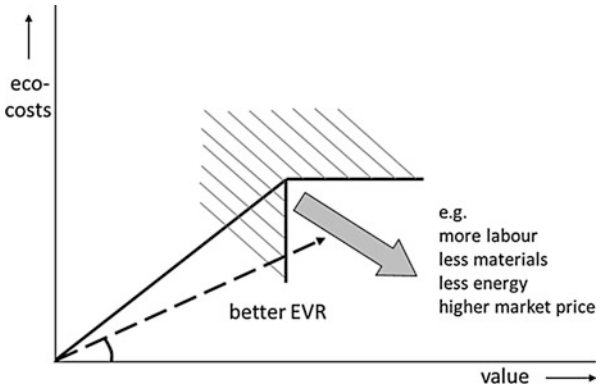



**Fig. 13** The required direction of "decoupling" ecology and economy: less eco-costs but more value (the double objective of Eco-efficient Value Creation)

### Eco-efficient Value Creation and the Double Objective

The conclusion of the analysis of the rebound effect is that sustainable products must have *lower eco-costs but at the same time higher value (market price)*. This "double objective" of designers in "Eco-efficient Value Creation" is depicted in Fig. 13.

Eco-efficient Value Creation and the double objective of Fig. 13 is not only to avoid the rebound effect: it deals also with the essence of the Brundtland report and the WBCSD mission statement as quoted in section "Introduction." It shows that care for the P of Planet and the P of Prosperity in the Triple-P model in section "Sustainability and Design for Sustainability Explained" is not an issue of trade-off (as proposed in (Elkington 1997)): it is not an issue of "or" but an issue of "and." The Brundtland report gives a vision on this issue (Brundtland 1987, page 6 of the summary): "Yet in the end, sustainable development is not a fixed state of harmony, but rather a process of change in which the exploitation of resources, the direction of investments, the orientation of technological development, and institutional change are made consistent with future as well as present needs." In the EVR model, the future needs are indicated by the eco-costs, the present needs are indicated by the value.

The reason we need to create extra value for eco-efficient products is threefold:

- People do not buy products with a lower value (compared to products with similar functionality and quality).
- A higher price in the market is required to cover the higher production costs of green products (note that a higher price is only accepted by the consumer when the perceived value is higher, otherwise the consumer will not buy the product).
- A higher price prevents the rebound effect.

It is the talent of the designer that creates the value of the product. It is EcoDesign in combination with LCA that helps the designer to reduce the eco-costs in a comprehensive way. Cases are given in section "Examples." A promising way of creating eco-efficient value is the design of Sustainable Product Service Systems, SusPSS.

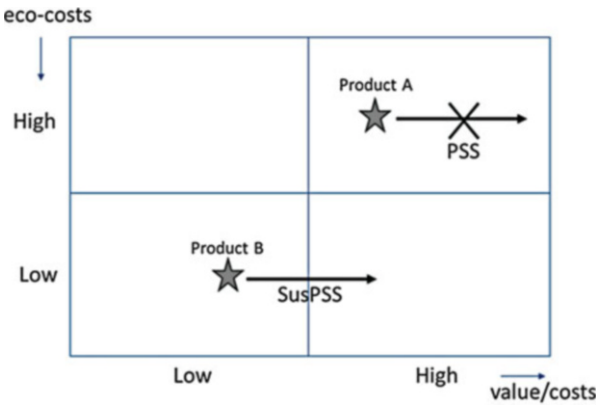



**Fig. 14** The effect of adding a service to a product. The PSS is non-sustainable. The SusPSS is sustainable

## The Sustainable Product Service System (SusPSS)

### Sustainable and Unsustainable Product Service Systems

The effect of adding a service to a product (creating a Product Service System, PSS) is explained in the Sustainable Business Strategy Matrix shown in Fig. 14.[4] Adding a service to a product results in an arrow to the right: the service increases the value and increases the eco-costs only with a very small amount (certainly, the eco-costs are not lowered by the extra service).

Product A, a relatively dirty product, in a PSS stays dirty as well. However, there is an added value because of the added service. The added value is used in business for two purposes:

- To enhance the profit margin (the costs of the service is less than the added value)
- To sell more of the product (which is normally the business aim of the PSS)

It is obvious that the business aim of selling more "dirty" products is not in line with sustainability. The product suffers from a high eco-burden, which cannot be lowered by adding a service. A comprehensive literature study on the subject underlines this general conclusion (Tukker 2004; Tukker et al. 2008).

Product B, a relatively clean product, however, suffers from a low relative value (which is often the case for green products). A designer may want to make the product more attractive to the market and may want to fulfill the double objective (creating lower eco-costs as well as higher value, compared to the reference product) by adding a service. This is the case of a Sustainable PSS, a SusPSS.

[4]Parts of this section have been copied from Chap. 6.2 of the book on Eco-efficient Value Creation (Vogtländer et al. 2013).

An example of the importance of the relative position of a product, and the use of PSS (Fig. 14), is a car-sharing system. Basically there are two groups of users of such a system:

1. People who went by bike and train before and want to have more convenience
2. People who owned a car before but do not want to invest in the next (second-hand) car since they do not drive much

It is obvious that the transfer "from bike to car" of the first group does not help our environment, especially not in the case of an internal combustion car. The second group "who had a car before" might change their behavior a bit, since they might drive less when they have to pay per km. However, the overall effect of the mix of the two groups is not expected to be positive for our environment (Meijkamp 2000).

The situation gets much better in the case of an electric car. The first group will pollute a bit more, however, much less than in the case of the internal combustion car. The second group will pollute less, since they shift from fuel to electric. This is a Sustainable PSS, a SusPSS.[5] The fact that some electric car-sharing providers offer an internal combustion station wagon for long-range holiday trips makes it even better: the added value of such an extra service outweighs the extra pollution for that period in terms of EVR.

Apparently, a PSS is only a SusPSS when it is applied to products which have a low score on the relative eco-costs. The eco-costs of dirty products can only be improved by a redesign of a product, not by adding a service. There are three basic ways to enhance the value of green products by creating a PSS:

1. Financing the product (postponing the investment)
2. Adding convenience
3. Adding image or fun (adding a "special experience")

### Concluding Remarks

When a product is dirty as such, PSS does not help to make the product cleaner. On the contrary, a PSS is unwanted in that situation, since it attracts extra buyers. The only right thing to do with a dirty product is to redesign it and improve its sustainability by reducing materials, energy, and transport, while maintaining a high value/cost ratio. When a product is green, and has a poor relative value, then the designer has to add quality and/or the company has to add a PSS (by upfront financing, adding convenience, enhancing image). See Fig. 15.

[5]Note that the situation for carpooling is completely different from car sharing: carpooling is always good for the environment, since it results in more passenger kilometer per car kilometer. Carpooling is an example of behavior in the use phase, rather than a PSS, since it is normally done between colleagues and friends (it is not a business as such).

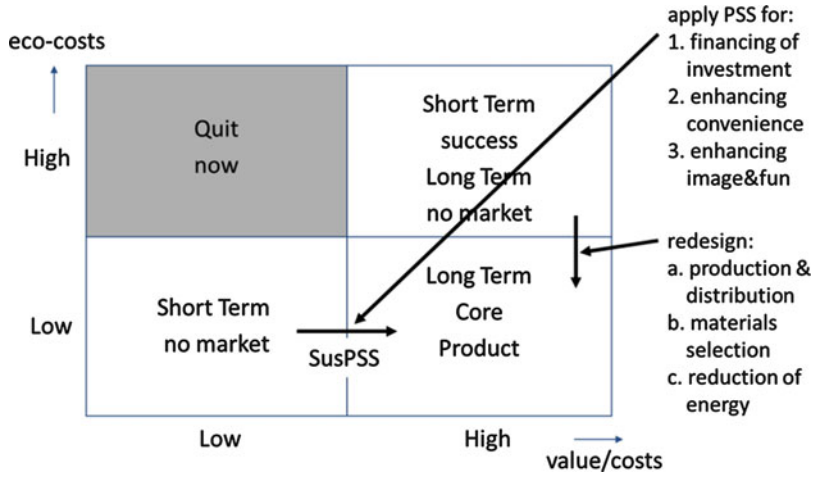



**Fig. 15** Sustainable enhancement strategies of the EVR of a Product Service System

## Examples

### Design of Luxurious Cork Products

Cork, a natural, recyclable, non-toxic, and renewable resource, which stems from the bark of a cork oak in the Mediterranean cork forest (Montado), is an optimal material for sustainable product design.[6] This section describes a project, developed for the Portuguese cork industry, on the sustainable innovation of cork products, using the method of "design intervention" combined with the method of Eco-efficient Value Creation (Mestre and Vogtländer 2013).

Design intervention is a method to generate innovative products in a structured way with a team of designers, focusing on maximum customer-perceived value. The method has four levels: the project strategic level, the concept development level, the design implementation level, and the product diffusion level. It includes workshops, combined with work in the design studios of the individual designers. The design concepts are analyzed with respect to sustainability, and the market value (WTP) of the prototypes is tested, see Fig. 16.

The results achieved with of the design method for four designs are depicted in Fig. 17 by means of a matrix similar to the product portfolio matrix as introduced in Fig. 10, yet now with the "relative value" = "value of the new design"/"value of the existing product" on the horizontal axis and with

[6]Parts of this section have been copied from Mestre and Vogtländer (2013). Details, figures, and tables can be found in this chapter.

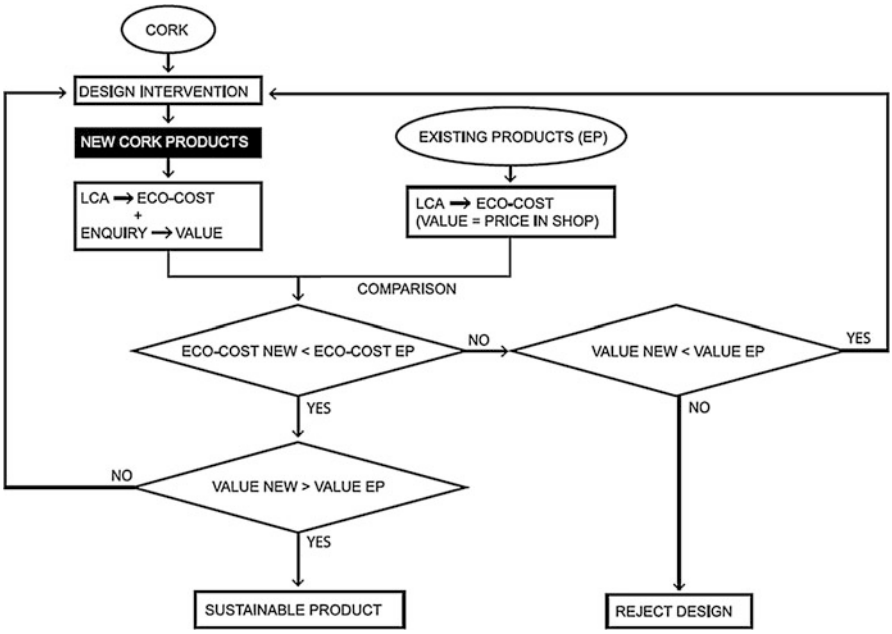



**Fig. 16** The process of Eco-efficient Value Creation of luxurious cork design products

"relative eco-costs" = "eco-costs new design"/"eco-costs existing product" on the vertical axis. The interpretation of the four quadrants in Fig. 17 is:

- Quadrant A: "quit" (since the eco-costs of the new product is higher and the value is lower than the existing product)
- Quadrant B: "reduce eco-costs"
- Quadrant C: "improve value"
- Quadrant D: "sustainable products"

One of the design cases in the project concerned the cork thermos flask. This thermos flask is a compact "espresso coffee thermos" which can be taken to the office, cinema, outdoor sports, or leisure activities. It is made of a combination of high-density agglomerated cork available in the Portuguese cork industry and the traditional Dutch porcelain from the Netherlands, thus exploring the functionality and the characteristics of the two traditional materials cork and porcelain, bringing together two ancient European technologies. It was designed by Tomas Schietecat and Boudewijn Van Limpt for Design Cork (Mestre 2008). The first design of the cork thermo flask was still a combination of cork and polypropylene, and the result (see point #18 in Fig. 17) is a good eco-cost, however, the value is low due to polypropylene solution for the cups. A redesign was considered with a substitution of the polypropylene for a higher value

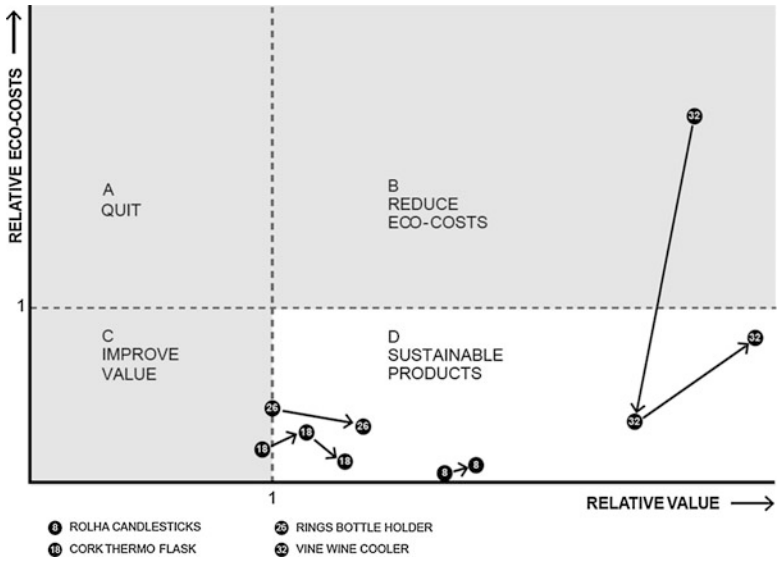



**Fig. 17** The process of Eco-efficient Value Creation depicted for four design cases

option polycarbonate, however, this second design was not chosen as the optimum in eco-costs and cost (point #19 in Fig. 17). The last redesign had an inner porcelain container, combining low eco-cost and high value (point #20 in Fig. 17), also due to use of Delft porcelain (a cultural high-end reference in European ceramics).

The project generated good results and that the matrix of Fig. 17 was a good means to depict this. These results were (see Mestre and Vogtländer 2013 for details): 27 out of 36 new designs ended up with better characteristics (lower eco-costs at a higher value) than their reference products; 7 designs were abandoned because of higher eco-costs than the reference products; 2 designs had low eco-costs, but were abandoned because of a low value; and about 50 % of the designs showed a factor 4 or more reduction of eco-costs, relative to the reference product. Products have been exhibited in several international cities, and 13 of them are sold in design shops in Lisbon, Porto, New York, Los Angeles, and Tokyo (website Corque Design 2013). The products are mostly related to interior design, such as furniture, and include such designs as "puff up" a string of cork beads that can be used to free-form a sitting object.

Note that because of the fact that all these luxurious products have an extremely low EVR (in comparison to, e.g., driving cars, living in houses, eating in restaurants), the focus of this type of sustainable product design must be on the value (WTP), in order to tempt consumers to spend their money on these products rather than on car driving or other activities with a high EVR (in line with the theory of the *indirect* rebound effect).

## Packaging Solutions for Food

The *classical sustainability perspective* on packaging is to reduce the environmental impact of the packaging, using life cycle assessment to evaluate different design alternatives. In this perspective, a brown paper bag (or even better: no bag) is the best solution.[7] Simultaneously, the *classical marketing perspective* on packaging is to generate value through differentiation, for instance through providing additional convenience. These two perspectives conflict, however, and the two-dimensional approach of the EVR model provides again a practical answer to the design dilemmas.

Of analyses of several types of packaging for food (Wever and Vogtländer 2012, 2013) we here give two examples:

1. Tomato ketchup bottles of glass and of PET.

   The LCA calculation on the eco-costs of the tomato ketchup bottles is given in (Wever and Vogtländer 2012), and the results are depicted in Fig. 18. This figure shows that the replacement of the glass bottle by a PET bottle is an instance of Eco-efficient Value Creation:
   - The eco-costs of the PET bottle system are lower than the glass bottle system, mainly because of the low weight of the PET bottle.
   - The value (i.e., the price in the retail shop) of the PET bottle is higher, partly because the PET bottle is squeezable, which is convenient in cases of high-viscosity products.
2. Water bottles with and without a sports cap.

   The LCA summary of the water bottles is given in (Wever et al. 2012). The results are depicted in Fig. 19. This figure shows that the redesign of the cap is not an example of the double objective, since the added value of a sports cap requires added materials, and therefore added eco-costs. It is what is expected in packaging innovations: the eco-costs are a bit higher. Yet, the EVR is lower because of the added value. Hence, the sports cap is supporting sustainability in terms of the theory on the *indirect* rebound effect and can still be regarded as an example of Eco-efficient Value Creation.

   As expected, closed loop recycling of the PET bottle system brings the innovation in the quadrant of the Eco-efficient Value Creation. Additionally, consumers might experience an increased inclination to reuse the bottle with the sports cap because of its enhanced relative value, diminishing the eco-costs by a factor 2 or more. Such mechanisms underline the importance of design as a value-adding activity for the relationship between environmental impacts, customer-perceived value, and consumer behavior. The sports cap design can achieve higher environmental gains on system level than the conventional design with the lowest eco-costs, because of its higher value.

[7] Parts of this section have been copied from (Wever and Vogtländer 2012). Details, figures, and tables can be found in this chapter.

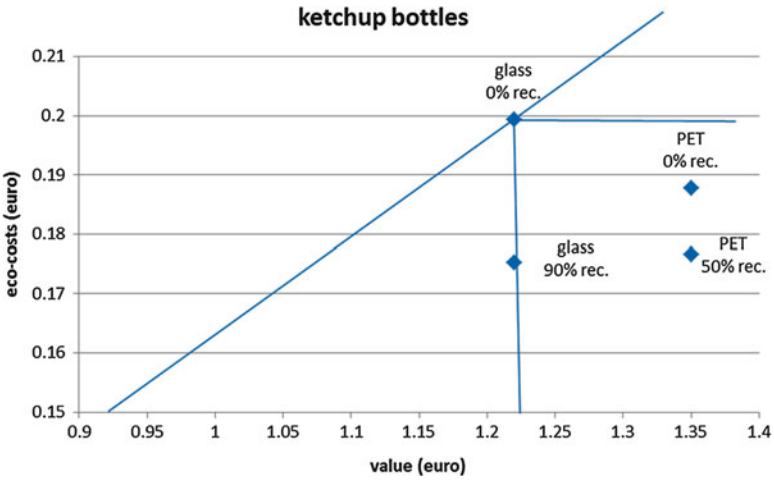



**Fig. 18** The value and the eco-costs of 300 ml tomato ketchup in a glass or PET bottle

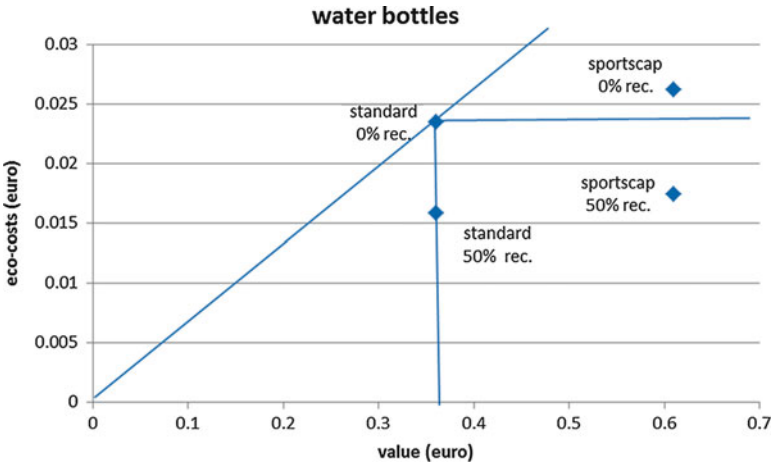



**Fig. 19** The value and the eco-costs of 50 cc water, standard or with sports cap

## Sustainable Water Tourism, an Example of a SusPSS

An interesting case of converting a PSS to a SusPSS is the case of sustainable water tourism in a lake district in the province of Friesland in the northern part of the Netherlands (Scheepens et al. 2014), with its main area the "Friese Meren" in the province of Friesland. See Fig. 20[8].

[8]Parts of this section have been copied from Scheepens et al. (2014). Details, figures, and tables can be found in this chapter.

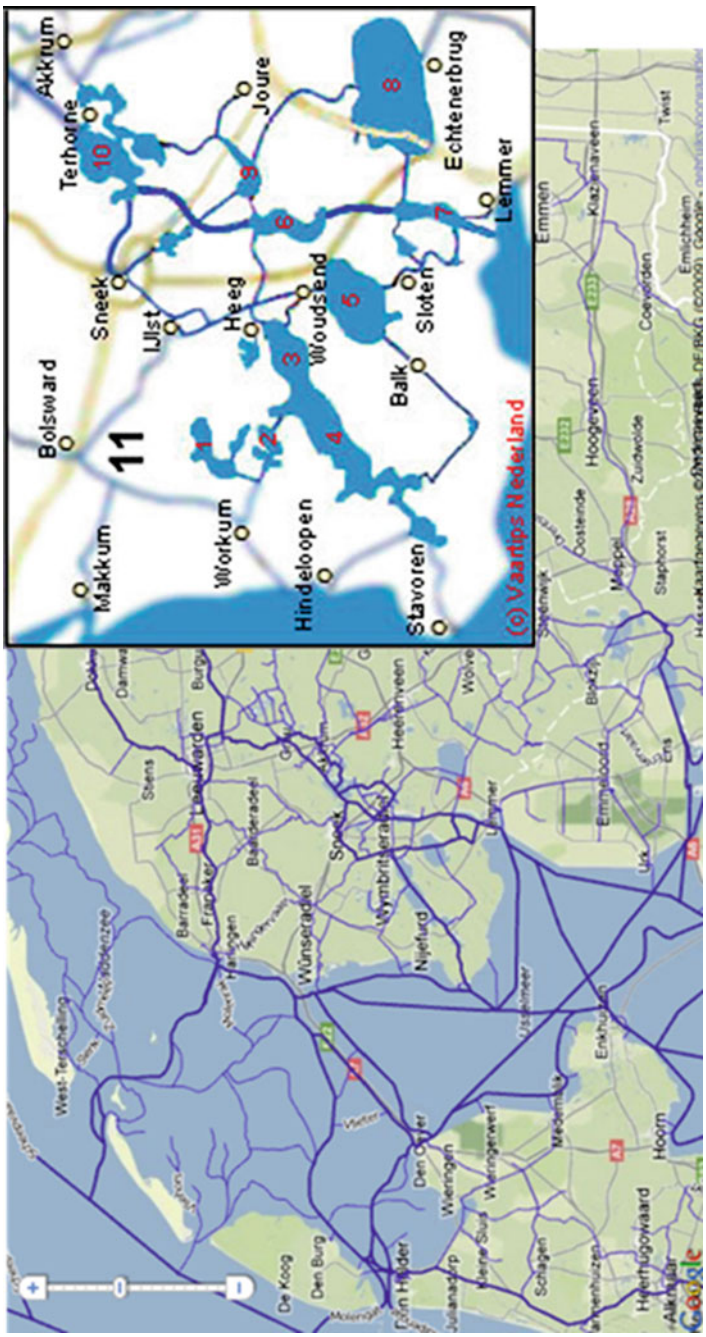



**Fig. 20** The province of Friesland and the Frisian Lake district

This area has had a stable number of tourists for the last decennium: approximately 1.2 million overnight stays in marinas per year. The question is, however, what can be done to expand the regional tourist industry and at the same time, reduce the regional pollution of the lakes, reed lands, and surrounding canals.

In a sense, this is the double objective in design at a regional scale:

- Higher value of the existing product service systems (e.g., rental of family boats in combination with the "experience" of the regional nature and the regional hospitality industry)
- Less regional pollution caused by these tourists

To achieve this double objective, the province of Friesland has started to following projects:

1. The development and introduction of a "water navigation system," which is an Internet information system on waterways, lakes, reed lands, and other natural areas of interest, social activities in villages, and advertisements of local shops, restaurants, museums, etc.
2. Subsidies for the introduction of a vast grid of charging points for electric vessels in marinas, in combination with sustainable energy sources
3. Subsidies for conversion of diesel propulsion systems of (rental) vessels to (hybrid) electric propulsion systems
4. Restriction of access to wet areas where nature has to be protected (electric boats are allowed, diesel and petrol boats are forbidden)

The ultimate goal is to trigger a total sustainable redesign of the vessels themselves.

The development of point 1 is, when it is used as a "stand-alone" PSS, an unsustainable development. It results in more tourists but also in more pollution (since they use conventional diesel boats).

The introduction of (hybrid) electrical boats of points 2 and 3 results in a drastic reduction of eco-burden. The costs of boats with electric propulsion systems are, however, approximately 20 % higher when compared to diesel and petrol vessels, which appears too much in the boat rental industry (as well as in the boat owners market). Giving subsidies is, moreover, not a sustainable solution for the total water recreation industry.

The trick for a successful introduction of this SusPSS is point 4. When diesel boats are forbidden in natural areas, this will create added value to the electric boats (it has already been proven at another Dutch lake area: the Nieuwkoopse Plassen), making subsidies superfluous.

So the bundle of measures fulfills the aforementioned double objective of Eco-efficient Value Creation. The results of the sustainable redesign, including *direct*, *indirect*, *economic-wide*, and *long-term* rebound effects, are depicted at Fig. 21.

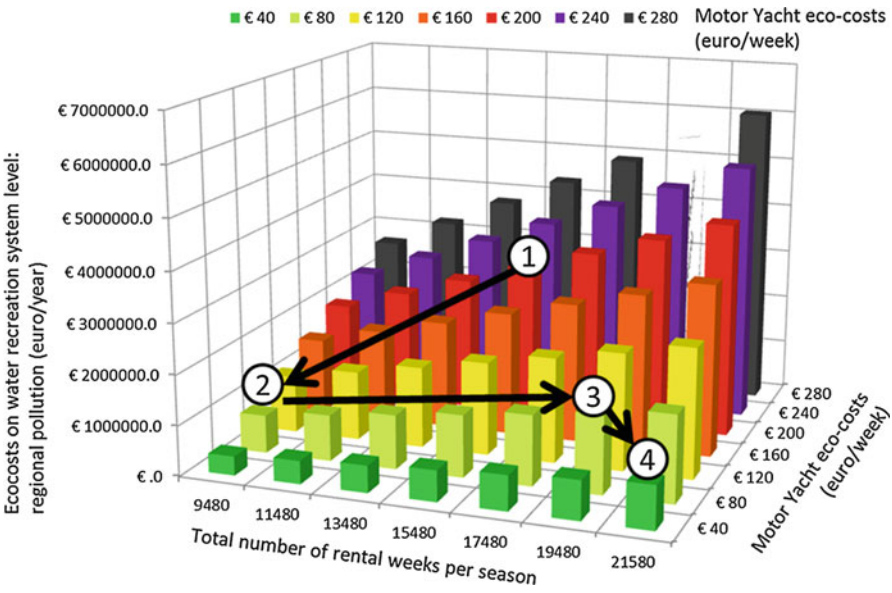



**Fig. 21** Tentative effect of the introduction of electric or hybrid boats, in combination with protected natural areas. *1* is the current situation, *2* is the potential decline of tourist by replacing all diesel boats by electrical boats, *3* is the potential growth of tourists by introducing large protected natural areas, and *4* is the introduction of special designed sustainable vessels

## Open Issues and Future Work

This chapter aimed at addressing design for the value of sustainability, with the understanding that sustainability is aimed at balancing economic, environmental, and social aspects (people, planet, profit). There is also design which is fully focused on either Planet (usually part of the art world, aimed at providing societal commentary) or on People (usually aiming to resolve some kind of social injustice). Both often heavily depend on public funding or fall in the higher "artistic" price segment. Although this type of design clearly has a substantial role to play in society and is often the trailblazer for economically viable projects, it is not aimed at reconciling all three pillars that are commonly deemed to constitute sustainability. Hence, they fall outside the scope of design for the value of sustainability.

Within the field of sustainability though, there are many schools of thought, very apt at disqualifying other approaches, e.g., cradle-to-cradle professionals criticizing eco-efficiency professionals or vice versa. In the end, all approaches may contribute, and all approaches have their shortcomings. Reconciling these approaches and different levels of ambitions is one of the major open challenges for Design for Sustainability.

Also developing a quantitative assessment for the social component of sustainability, like the environmental ones presented in section "Life Cycle Assessment ("Fast Track")" on Life Cycle Assessment, is a long-held wish of the sustainable design community.

In search of the ultimate tool for designers, there is a debate on why the vast majority of designers struggle with the application of the methods and tools described in this chapter (Vallet et al. 2013; Lofthouse 2006). The major complaints of designers are that all forms of design for the value of sustainability are too complex and too laborious, and they do not fit into their design practice. Designers want tools which are simple and at the same time detailed to answer complex questions, and they want to have easy access to examples which help them resolving their specific answers. Both sets of requirements seem to be contradictory in itself, but the need for inspiration and information is evident. The likely solution is to add information on sustainability to sector-specific computer-aided design software (e.g., ArchiCAD for architects, SolidWorks for industrial designers) rather than develop comprehensive stand-alone software for sustainable design.

Relative outsiders, dealing with Design for Sustainability, should be aware of the different approaches and strive to align their organization's intentions with the appropriate approach and appropriate designers. It should be noted here that there is the complication that not all designers talking about the same approach will use identical terminology nor do designers using identical terminology necessarily mean the same thing. Furthermore, the correct application and/or strict adherence to an approach is not a given. In that respect, consolidation of terminology and standardization of approaches is a final open issue.

## Conclusions

Design for Sustainability can be described as an attempt to include the issue of sustainability in the design process. Sustainability is then taken as an extra design criterion, added to all other design aspects and design criteria.

Approaches to Design for Sustainability come in three groups:

- The holistic approaches of cradle-to-cradle, Circular Economy, and Biomimicry (and many other related approaches).
- The checklist (environmental benchmarking) and tools approach of EcoDesign, Design for the Environment, Design for Sustainability, the LiDS Wheel, and at a more detailed level Design for Recycling and Design for Disassembly.
- The quantitative approach of Life Cycle Assessment (LCA) and the LCA-based model of the Eco-costs/Value Ratio (EVR).

The challenges connected to the holistic approaches are on how to make them usable in day-to-day design. The challenges to the checklist approach lie in the question of sufficient improvement to truly be effective. The challenges around the

quantitative modeling are around making the correct comparison between alternatives and in choosing what to include and what to exclude from a study.

The different approaches in the three groups mentioned above are complementary and should be applied in all design stages. The challenge for relative outsiders, dealing with Design for Sustainability, lies in selecting the most appropriate combination of approach and designer for a given problem.

## Cross-References


- ▶ Design for Values in Engineering
- ▶ Design Methods in Design for Values


## References


Bakker (2012) Design for sustainable behaviour; three approaches. Research Lecture, Design Engineering Department, Delft University of Technology, 19th Apr 2012

Benyus JM (1997) Biomimicry: innovation inspired by nature. HarperCollins, New York

Bhamra T, Lofthouse V (2003) Information/inspiration: a web based sustainable design tool. Repository of the Loughborough University. See also http://ecodesign.lboro.ac.uk/index.php?section=72

Biomimicry (2013) Biomimicry 3.8 http://biomimicry.net/ last seen Oct. 6 2014

Boothroyd G, Dewhurst P, Knight W (2002) Product design for manufacture and assembly, 2nd edn. Marcel Dekker, New York

Boulding KE (1966) The economics of the coming spaceship earth, from Environmental quality in a growing economy. pp 3–14. Available at www.eoearth.org/article/The_Economics_of_the_Coming_Spaceship_Earth_%28historical%29. Accessed 05 Oct 2013

Brezet H, Van Hemel C (1997) Ecodesign: a promising approach to sustainable production and consumption. UNEP, Paris

Brundtland GH (1987) Our common future. United Nations World Commission on Environment and Development. Oxford University Press, Oxford

BSI, British Standards Institution (2011) PAS 2050: 2011 specification for the assessment of the life cycle greenhouse gas emissions of goods and services. BSI, London

Chiodo JD, Grey C, Jones D (2011) Design for remanufacture, recycling and reuse, ICOR. 27–29 July 2011, Glasgow. See also www.activedisassembly.com

Corque Design (2013) Corque Design http://www.corquedesign.com/ last seen Oct 6 2014

Crul MRM, Diehl JC (2006) Design for sustainability, a practical approach for developing economies. UNEP, Paris

Crul MRM, Diehl JC, Ryan C (2009) Design for sustainability, a step by step approach. UNEP, Paris

Daae J, Boks C (2013) Dimensions of behaviour change; designing how users interact with products [a design tool consisting of card set]. NTNU, Trondheim

ecocostsvalue (2013) The model of Eco-costs / Value Ratio (EVR). http://www.ecocostsvalue.com/ last seen Oct. 6 2014

Elkington J (1997) Cannibals with forks. New Society Publishers, Gabriola

EN 15804 (2012) Sustainability of construction works. Environmental product declarations. Core rules for the product category of construction products (EN 15804:2012). ISO/FDIS, Geneva

European Commission (2012) "Manifesto for a resource efficient Europe" MEMO/12/989. Available at http://europa.eu/rapid/press-release_MEMO-12-989_en.htm. Accessed 05 Oct 2013

Gosseries A (2008) Theories of intergenerational justice: a synopsis. SAPIENS 1(1):61–71
Guinee JB (ed) (2002) Handbook on life cycle assessment, operational guide to the ISO standards. Kluwer, Dordrecht
ILCD (European Commission, Joint Research Centre, Institute for Environment and Sustainability); International Reference Life Cycle Data System (ILCD) (2010) Handbook: general guide for life cycle assessment (LCA) – detailed guidance, First edition. Free available on www.lct.jrc.ec.europa.eu/publications
ISO 14040 (2006) Environmental management – life cycle assessment – principles and framework (ISO 14040: 2006). ISO/FDIS, Geneva
ISO 14044 (2006) Environmental management – life cycle assessment – principles and framework (ISO 14044:2006). ISO/FDIS, Geneva
ISO 14067 (2013) Greenhouse gases – carbon footprint of products – requirements and guidelines for quantification and communication (ISO/TS 14067:2013). ISO/FDIS, Geneva
JRC (2013) LCA tools, services and data. http://eplca.jrc.ec.europa.eu/ResourceDirectory/services2-1.vm; last seen Oct. 6, 2014.
Lilley D (2009) Design for sustainable behaviour: strategies and perceptions. Des Stud 30(6):704–720
Lilley D, Wilson G (2013) Integrating ethics into design for sustainable behaviour. J Des Res 11(3):278–299
Lockton D, Harrison D, Stanton N (2008) Making the user more efficient: design for sustainable behaviour. Int J Sustain Eng 1(1):3–8
Lockton D, Harrison D, Stanton NA (2010) The design with intent method: a design tool for influencing user behaviour. Appl Ergon 41(3):382–392
Lofthouse V (2006) Ecodesign tools for designers: defining the requirements. J Clean Prod 14(15–16):1386–1395
MBDC (2013) C2C Framework. http://www.mbdc.com/cradle-to-cradle/c2c-framework/last seen Oct. 6 2014
McDonough W, Braungart M (2002) Cradle to cradle, remaking the way we make things. North Point Press, New York
McKinsey (2012) Towards the circular economy, vol 1. Ellen MacArthur Foundation, Isle of Wright, UK
McKinsey (2013) Towards the circular economy, vol 2. Ellen MacArthur Foundation, Isle of Wright, UK
Meijkamp RG (2000) Changing consumer behaviour through Eco-efficient Services - An empirical study on Car Sharing in the Netherlands. Doctoral dissertation, Delft University of Technology
Mestre A (2008) Design Cork for future, innovation and sustainability. Lisbon: Susdesign
Mestre A, Vogtländer J (2013) Eco-efficient value creation of cork products: an LCA-based method for design intervention. J Clean Prod 57(15):101–114
Petala E, Wever R, Dutilh C, Brezet H (2010) The role of new product development briefs in implementing sustainability; a case study. J Eng Technol Manag 27(3/4):172–182
Prahalad CK (2002) The fortune at the bottom of the pyramid. Pearson Prentice Hall, Upper Saddle River/New York
Recoup (2009) Plastics packaging. Recyclability by design. RECycling of Used Plastics Limited, Peterborough
Scheepens AE, Vogtlander JG, Brezet JC (2014) Two LCA based models to analyse complex (regional) circular economy systems. Case: making water tourism more sustainable. J Clean Prod, submitted
Sorrell S (2007) The rebound effect: an assessment of the evidence for economy-wide energy savings from improved energy efficiency. UK Energy Research Centre, London
Stahel W (1982) The product-life factor. http://infohouse.p2ric.org/ref/33/32217.pdf and http://product-life.org/en/major-publications/the-product-life-factor. Both accessed 05 Oct 2013
Tobin J (1974) What is permanent endowment income? Am Econ Rev 46(2):427–432

Tukker A (2004) Eight types of product – service systems: eight ways to sustainability? Experiences from SUSPRONET. Bus Strat Environ 13:246–260
Tukker A, Charter M, Vezzoli C, Stø E, Munch AM (2008) System innovation for sustainability, vol 1. Greenleaf Publishing, Sheffield
Vallet F, Eynard B, Millet D, Mahut SG, Tyl B, Bertoluci G (2013) Using eco-design tools: an overview of experts' practices. Des Stud 34(3):345–377
Van Boeijen A, Daalhuizen J (2013) Delft design guide. BIS Publishers, Amsterdam
Van Hemel CG (1998) Ecodesign empirically explored. PhD thesis. Repository of Delft University of Technology
Vogtländer JG (2012) A practical guide to LCA for students, designers and business managers, 2nd edn. Delft Academic Press, Delft
Vogtländer J, Mestre A, Van der Helm R, Scheepens A, Wever R (2013) Eco-efficient value creation, sustainable design and business strategies. Delft Academic Press, Delft
WBCSD (1995) Achieving eco-efficiency in business, Second Antwerp eco-efficiency workshop, 14–15 Mar. World Business Council for Sustainable Development
Wever R (2012) Design research for sustainable behaviour. J Des Res (Special Issue) 10(1):1–139
Wever R, Vogtländer J (2012) Eco-efficient value creation: an alternative perspective on packaging and sustainability. Packag Technol Sci 26(4):229–248
Wever R, Vogtländer J (2013) Assessing the relative sustainability of different packaging sizes. In: 26th IAPRI symposium on packaging, Helsinki, 10–13 June 2013

# Design for the Value of Trust

Philip J. Nickel

## Contents



**Abstract**

The relationship between design and trust has recently been a topic of considerable scholarly discussion. This is due to several reasons. First, interpersonal trust is an especially relevant concept in information, communication, and networking technologies, because these technologies are designed to facilitate transactions and exchanges between people. Second, digital information has become ubiquitous and can itself be the object of a trust-like attitude, since people rely on it to meet their expectations under conditions of time and information scarcity. And finally, perhaps as a result of the first two points,

Research for this article was partly supported by the MVI (Responsible Innovation) program of the Dutch NWO (Netherlands Organisation for Scientific Research), within the project "Medical Trust Beyond Clinical Walls."

P.J. Nickel (✉)
Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: p.j.nickel@tue.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_21

designers have started to take on the role of expressly encouraging user trust by incorporating in their designs perceptual and social cues known to increase trust. This chapter explores some of the philosophical issues surrounding trust "by design" and explains how to apply Design for Values to trust.

**Keywords**
Trust in technology • Technological mediation • Trustworthiness • Ethics of trust • Epistemology of trust

## Introduction

A traditional approach to design for trust is to make artifacts, processes, and systems that are reliable or trustworthy. A car is reliable when it has been designed to function safely and efficiently, and this reliability fosters trust in the technology and the company that produces it. Recently, however, designers have taken on a different role, inviting trust directly by using perceptual and social cues known to encourage trust (e.g., Glass et al. 2008). The focus then shifts from the reliability of the system to the psychological state of the user. Whereas previously the psychological state of the user was left to the user himself or to the advertising department, now it is a focus of design. This shift in focus was caused in part by the ICT (information and communications technology) revolution, in which technologies were designed to mediate social relationships. The dispositions of users to place trust in other users became an explicit subject of design (Riegelsberger et al. 2003). As a result, design for trust in its new incarnation is *reflexive*: it encompasses both the creation of reliable and trustworthy products and systems and also explicit reflection on the trust of the user.

Trust is a good thing, but for many years the academic literature on trust has emphasized that it is not good always and everywhere (Baier 1986; Coleman 1990; Hardin 1993). The reason is not just that the value of trust is sometimes overridden by other values such as security or fairness. After all, many important values can be overridden in some circumstances by other more salient values. The reason is more pointed: unlike some other values, trust is a psychological state that represents the trusted person or object as being trust*worthy*, and this may or may not actually be the case. When the thing one relies on is not trustworthy, then trust is inappropriate or even dangerous. In other words, trust involves accepting one's vulnerability to others, willingly placing oneself in their hands to some extent. Therefore, to encourage trust is to encourage a kind of vulnerability (Baier 1986). This is why it is important for designers' current reflexive concerns about trust to remain coupled with a traditional concern for trustworthiness and reliability. In this chapter, I will not focus on reliability, responsibility, and the like, which provide backing or grounding to the value of trust, because these are separate values, treated elsewhere in this volume. I will instead keep the emphasis on the psychological state of trust in the user and the ethics of designing with this psychological state in mind, so as to deal with what is distinctive about trust.

This chapter introduces some prevalent conceptions of trust and indicates how they can be used to inform design. In the section "The Value of Trust", the definition of trust is set out and some disagreements about how to conceptualize it are discussed. In the section "Design for Trust", several philosophical approaches to the idea of design for trust are discussed, and the methodology of Design for Values is applied to the value of trust. In the section "Cases and Examples", a number of case studies of design for the value of trust are described. The sections "Open Issues and Future Work" and "Conclusions" raise questions for future work and conclude the chapter, respectively.

## The Value of Trust

### Conceptions of Trust and Its Value

In order to design for trust, we want to know both what trust is and when it is appropriate or inappropriate. The first question – "What is trust?" – depends on what we hope to explain using the concept. Social scientists and philosophers agree, by and large, that trust refers to human reliance that is:

- Willing and voluntary
- Carried out under conditions of uncertainty and vulnerability

Trust is often conceptualized as a relationship between a trustor, a trustee, and a desired performance or a domain of interaction or cooperation. This is called "three-place trust" (Baier 1986), in which one person trusts a second person or entity to do some particular thing (e.g., to make a payment) or to promote his or her interests in a domain of shared interaction (e.g., the financial domain). Thinking in this way, we can focus on certain *performances* by the trustee that fulfill or satisfy the trustor's trust-based expectations.

Although social scientists and philosophers are also sometimes interested in "two-place" trust, in which a person simply trusts another person or entity, with no particular performance or domain in mind, this two-place notion of trust is usually understood as being derivative from and less explanatorily useful than three-place trust. For this reason it will not be emphasized in what follows. It can, however, be fruitful to consider a less conscious notion of "basal" trust (the term comes from Jones 2004) that concerns the regular behavior of the natural world, the functioning of one's own body and faculties, of social practices and institutions, and of the built and engineered environment. Throughout our lives we rely on tacit assumptions about how things will or are "supposed to" behave (Carusi 2009). This basal trust only becomes visible when there is some kind of breakdown (Jones 2004) or when we imagine a scenario of breakdown (Nickel 2010). Technology can induce this as well (Carusi 2009). Despite its implicitness, basal trust is not the same thing as two-place trust. Arguably, even in basal trust a person trusts some entity to do something, even if this reliance is tacit and unreflective.

In addition to questions about what trust is, there are also evaluative questions concerning when it is appropriate or good to trust and concerning the importance of trust in relation to other values. Suppose you are using a Web-based tool for project collaboration and want to know whether you can rely on another user whom you do not know personally. Should the system give you *reasons* to believe that the person is trustworthy before you are expected to interact with him or her, or would that impose an unrealistic barrier that gets in the way of the practical value of cooperation? Emphasizing the evidential dimension of trust over its practical and pragmatic dimension, or vice versa, can yield different outcomes for design. In addition, what kinds of other values within design strengthen or weaken trust? For example, accessibility by many users might be "democratic," but it may also introduce untrustworthy users, thereby decreasing trust. Rigorous security and safety measures, on the other hand, seem to take the place of trust rather than encouraging it.

## Issues of Controversy Regarding Trust and Its Value

There are a number of scientific controversies about the nature of trust, especially about the characteristic kinds of motives and evidence that underlie trust. First, philosophers tend to think of trust as a moral concept and say that moral or richly affective motivations underlie trust (McLeod 2002; Simpson 2011; Lagerspetz and Hertzberg 2013), whereas social scientists often leave this open or tie it to a nonmoral, impersonal motivation such as expectation that one will have additional interactions with the trusted party in the future (Coleman 1990; Hardin 2006).[1] Second, some stress that the idea of trust applies most clearly to people who know each other well and have an ongoing relationship, whereas others emphasize the importance of cooperation between strangers. For our purposes here, we will leave these questions open and consider trust in a broad sense including all these motivations.

However, some philosophical problems and controversies surrounding trust have a special relevance for design and will reappear in other parts of this chapter:

- Anthropocentrism: People often speak about trust in technology or in specific artifacts, but it seems inappropriate to take a rich affective or moral attitude toward a mere thing, rather than a person (Nickel et al. 2010) because this is a kind of pathetic fallacy, anthropomorphizing an object.
- Evidence for Trust: *Evidence* is highly relevant when figuring out whether to rely on another person or entity (e.g., a computer system or another system user), particularly when the stakes are high. There is disagreement about whether trust is typically based on evidence that the person or thing relied upon is reliable (Gambetta 1988) or whether it is a "leap of faith" carried out under conditions of uncertainty on the basis of non-evidential information (Möllering 2006). People

[1]Cf. Uslaner (2002) who links trust to a general moral worldview linked with personality and childhood experiences.

often trust on the basis of weak evidence, and so long as the trusted entity is actually trustworthy, it is not clear why having more evidence is better. Indeed, a policy of never trusting without conclusive evidence is usually harmful overall (Hardin 1993).

- Discretion of the Trusted: Greater assurance of reliability seems to increase trust, but on the other hand, constant surveillance and strict enforcement of performance (e.g., with legal sanctions) tend to make trust irrelevant (O'Neill 2002; Smolkin 2008). Trust is most relevant when the person or thing trusted has the *discretion* to choose how to behave.

These problems raise a number of questions about willing reliance upon a technology, and they prompt a closer look at the motives for trust. Psychological research has mostly shown that risk adversity and overall trustfulness are not closely related (e.g., Ben-Ner and Halldorsson 2010). This points toward an important conceptual truth about trust: although we may wish to say that trust is based on evidence, it is a different kind of evidence than that relevant to a risk judgment. Whereas risk judgments are based on predictions and associated emotional states such as fear, the evidence relevant to trust concerns something else entirely, namely, knowledge of others' motives and interests, norms regarding roles and relationships, and situational knowledge. Once one has a standing relationship with somebody on the basis of which normative expectations are formed, further evidence is no longer needed to trust them unless specific doubts arise. In addition, constant surveillance and strict enforcement make motive-based knowledge irrelevant. This means that how a technology mediates the motives of its users is a central issue of design for trust.

## Design for Trust

### Existing Approaches and Tools

This section looks at existing conceptions of design for trust and its methodology. Broadly speaking, we distinguish between anthropocentric conceptions of design for trust, which limit trust to interpersonal relationships, and non-anthropocentric conceptions which allow trust to have things other than humans as its object. Friedman, Kahn, and Howe (2000) define trust in a way that explains their anthropocentric approach: in trust one ascribes goodwill to others, allowing oneself to be vulnerable to them (citing Baier 1986). Goodwill requires consciousness and agency. But since "technological artifacts have not yet been produced … that warrant in any stringent sense the attribution of consciousness or agency," it follows that "people trust people, not technology" (Friedman et al. 2000, p. 36). Despite their focus on interpersonal trust, they are nonetheless especially concerned about trust with regard to the Internet and other ICTs. The reason is that these are technologies that facilitate social transactions and commercial exchange – situations where interpersonal trust between users is required or highly instrumental.

In the previous section, we raised the question of what kind of evidence should be made available to users in order to establish trust (the issue of *Evidence for Trust*). Friedman, Kahn, and Howe define the relevant Evidence for Trust in terms of three types of information:

- About possible harms
- About the motives of the persons with whom a user interacts by way of the technology
- About whether those persons' motives could cause the indicated harms (2000, p. 35)

In accordance with this view, Friedman, Kahn, and Howe often restrict design for trust to voluntary human factors in interpersonal interaction. They argue that it is important "not to conflate trust with other important aspects of social interaction" that could also fail, such as having insufficient information about an entity or person on which one relies (2000, p. 37). Harm that occurs "outside the parameters of the trust relationship" does not count against trust (2000, p. 35). Friedman, Kahn, and Howe put forward a number of engineerable factors that help cultivate trust online, such as reliability and security of the technology, protection of privacy, self-assessment of reliability, honest informational cues, accountability measures, and informed consent. According to them, these factors help facilitate interpersonal trust between buyers and merchants and between individuals who participate in online fora.

By contrast, others writing about design for trust have an inclusive, non-anthropocentric conception of trust. They include technology itself as an appropriate object of trust. For example, Kelton, Fleischmann, and Wallace argue that digital information is itself a paradigmatic object of trust, pointing out that "the overwhelming volume of information on the Internet creates exactly the type of complexity that gives rise to the need for trust" (2008, p. 368). They argue that some important hallmarks of trust are present in our relation to information on the Internet, namely, uncertainty, vulnerability, and dependence. A similar but broader argument could be given regarding our more general reliance on technological systems. Given that we are uncertain about, vulnerable to, and dependent on technological systems and that this is ineliminable because of the sheer complexity of these systems and their integral involvement in our daily lives, we should also be willing to speak of trust in technological systems while acknowledging that this is of a different nature than interpersonal trust (Nickel 2013). Our expectations of technology are not just predictive but also normative – they involve the attitude that the technology should perform in certain ways and should promote or protect our relevant interests. In that case, design for trust should provide cues and evidence that help people ground their trust in technological systems (*Evidence for Trust*). Design for trust implies that the designer pays attention to the user's expectations of a technological artifact and tries to create a condition in which the user has warranted expectations with regard to the actual functions of the artifact (Nickel 2011).

Another important strand of non-anthropocentric thought about design for trust draws on continental philosophy of technology. Heidegger, for example, has been

interpreted by Kiran and Verbeek (2010) as holding that technology can itself be an object of trust or suspicion. In this line of thought, trust signals a constructive relationship to technology. As users of technology, we do not take technology as a mere instrument for prior goals. Instead, we engage with it actively and takes responsibility for how our "existence is impacted" by it (2010, p. 424). The analysis of how technology *mediates* human action and freedom in this way, and of its foreseen and unforeseen effects on our concepts and social and ethical practices, provides a starting point for ethical and practical reflection. For example, Verbeek (2008) analyzes how imaging technology mediates medical decision-making, arguing that it has a profound impact on agency in this area. Although Verbeek (2008) does not mention trust, we can read Kiran and Verbeek's analysis of trust in technology back into the case of medical imaging in two ways. First, imaging technology complicates trust between physician and patient, a relationship in which trust is acknowledged to be of central importance (see, e.g., Eyal 2012). The trustworthiness of the physician and the reliability of the imaging technology are defined in relation to one another: the physician vouches for its reliability, and its detailed images may invite or require expert interpretation. But furthermore, trust by both the physician and the patient in the imaging technology is a kind of purposive, constructive engagement with technology and plays a powerful role in determining what is *going on* in a given encounter: what kinds of considerations are taken as relevant, how clinical consultations proceed, and what responses are considered standard. An analysis of trust in technology in this deeper sense can be useful for design since it gives insight into how a technology transforms or is likely to transform, our actions, perceptions, and practices. "Basal trust" is often the focus here, since basal trust concerns what we regard as "comfortable," "everyday," or "normal," creating the background assumptions framing human action and interaction.

Now that we have discussed both anthropocentric and non-anthropocentric accounts of design for trust, we can discuss how design methods incorporate the value of trust. How can value-based reflection discover that trust is important to a design, and how can this reflection be made a part of the subsequent design process? Design for Values, as described in other chapters in this volume, covers several methods for doing this. Here we will focus on value-sensitive design. Value-sensitive design is a specific approach to Design for Values that has been applied to such widely divergent cases as the user interfaces for missile-guidance systems (Cummings 2006) and office interior design (Friedman et al. 2006). As originally set out in Friedman et al. (2006), value-sensitive design consists of three related aspects or phases of research and design work: a *conceptual* phase in which relevant values and potential conflicts between them are identified, an *empirical* evaluation of how (well) various values are realized in various permutations of a design, and a *technical* phase that attempts to resolve conflicts between values or achieve a more effective realization of those values through engineered solutions. Here I will briefly discuss these three phases in relation to trust.

*Conceptual*. According to value-sensitive design, trust will sometimes emerge as an important value during the conceptual phase. But it is not always clear how this is

supposed to be determined: what process or criterion should be used? Manders-Huits, for example, raises the question of what counts as a value within value-sensitive design (2011). We can answer this question by referring to the conception of trust discussed above. There are two main cues that indicate that trust is a salient value in design. The first is whether a design mediates interpersonal relationships in a new way, or for new users, or in a new context, or whether existing relationship-mediating features of a technology are believed to be lacking in some way. The second is whether technology adoption occurs under conditions of uncertainty, dependence, and resource limitations for users, where the context is new or existing solutions are believed to be lacking. These two tests can be used to discover whether trust is a salient value in the conceptual phase of value-sensitive design or other Design for Values methodologies. This can be facilitated by explicitly discussing issues of trust during interactions with users and stakeholders, but often it will emerge naturally as a design problem is made more concrete.

*Technical*. When trust is discovered to be a key value, other design methods can be used to help determine how to balance it with other values and realize it technically within the design. Which method is used will depend on whether it is a redesign or a more radical innovation. Vermaas et al. (2010) embed design for trust within two existing design methodologies: *Quality Function Deployment* (King 1989; Akao 1990), used for redesign, and a creative design methodology for new designs as described by Cross (2006). In the case of redesign, "the values derived from trust . . . are listed as user requirements and . . . IT developers analyze which of the characteristics of their existing systems are relevant to meeting these values" (Vermaas et al. 2010, p. 502). In the case of creative design, an ongoing process of discussion results in an open dialogue in which "a 'space' of design solutions co-evolves with a 'space' of design problems" and strong engagement with users and clients is needed throughout the process as solutions evolve (Vermaas et al. 2010). Vermaas et al. view the use of these design methodologies as compatible with value-sensitive design, although they do not explicitly relate them to the structure of that approach which divides design tasks into conceptual, empirical, and technical phases.

*Empirical*. The empirical phase measures the realization of trust within different technical implementations of a design. Here it is important to remember three crucial points. First, trust is more than mere reliance. At a minimum, it is a voluntary disposition toward reliance under conditions of uncertainty. Trust uses information about the motives, interests, and character of individuals or about the functions of artifacts and systems, together with situational knowledge, to overcome uncertainty. For that reason, trust cannot be equated simply with a risk estimate (since estimating risk directly is not the characteristic basis of trust), nor can it be equated with a disposition to cooperate or engage in reliant behavior (since such a disposition might not be fully voluntary, e.g., when there are no other good options). If the expected behavior is certain, e.g., because it is being enforced coercively, then trust is not the explanation for one's reliance on that behavior. (This is the issue of the *Discretion of the Trusted* mentioned earlier.) Second, trust is usually thought to involve a normative expectation that somebody or something

*should* perform a certain way, or to a certain standard, and this normative expectation forms part of the reason for relying on the trusted entity. This is what distinguishes trust from a willing disposition toward reliance grounded in a purely predictive or statistical expectation. And third, trust is not the same thing as trustworthiness. When we measure trust, we are dealing with a psychological disposition centrally consisting of people's expectations. Trustworthiness, on the other hand, is a quality of a person, system, or artifact such that it is likely to perform as expected. Showing that people trust (within) a design does not imply that it is trustworthy, nor the other way around.

## Comparison and Critical Evaluation

It can be difficult to settle differences between philosophical views about design for trust, because each view emphasizes different elements that are important for design. Friedman, Kahn, and Howe's emphasis on trust as involving the willful actions of those who interact online, such as buyers and sellers, is supported by the philosophical literature emphasizing interpersonal trust. However, it sharply restricts the domain of design for trust. For example, some harms or bad outcomes incurred through acts of reliance are caused by technical problems or human accidents rather than the (ill) will of those interacting, and yet these technical problems and accidents also seem to affect trust. Suppose a Web merchant accidentally doubles an order and overcharges the buyer. It would seem to follow from Friedman, Kahn, and Howe's view that such an incident does not have to do with trust so long as the humans involved are non-culpable and non-negligent. Responsibility for the incident cannot be attributed to the merchant's ill will, after all. Since Friedman, Kahn, and Howe's guidelines for design for trust are intended to address cases such as these, they have to make an indirect argument, claiming that engineered factors affect trust because "people frequently draw on cues from the [engineered] environment to ascertain the nature of their own vulnerabilities and the good will of others" (Friedman et al. 2000, p. 37). However, this claim seems to assume that the technology has a strong mediating role in the formation and presentation of human motives. It does not sit easily with their claim, quoted earlier, that we must not confuse failure of trust with having insufficient (or incorrect) information about an entity or person on which one relies.

Broader, non-anthropocentric views of trust that allow for trust in technological artifacts do not have this problem, but they encounter the criticism that trust in technology is indistinguishable from mere reliance or judgments of reliability (Nickel et al. 2010). To some extent, this criticism can be met by giving an account of the moral, affective, or emotional aspects of trust in artifacts and technological systems, such as the frustration one feels when an artifact or system breaks down. Such emotions and normative judgments go beyond mere reliance. However, it may seem that normative, affective, or emotional attitudes about technological artifacts are irrational, since technologies are in the final reckoning just "brute matter." One may get angry at one's car when it does not start, but perhaps this does

not signal a rich relationship of trust yielding insights for design (although the concept of "reactance" to technology is taken seriously by human-technology interaction theorists – see, e.g., Lee and Lee 2009; Roubroeks et al. 2011). It has even been argued that affective, social attitudes toward computers, robots, or persuasive technologies should be discouraged by design and that encouraging these attitudes is morally questionable because it deceives technology users (Friedman 1995).

A compromise view could be reached by focusing on the situation of users who in various situations want or need to rely on technological artifacts and systems and have little time or expertise for the evaluation of whether this reliance is a good idea. Think for a moment of users who are bombarded with information and opportunities to use technologies and who have a finite supply of attention and cognitive resources to spend on the question of whether and how to use them. Various possible ultimate objects of trust such as other users, system operators, designers, manufacturers, owners, and the technology itself are not clearly separated in the user's mind. From a design point of view, it is more important to consider the parameters of user choices (How much time does the user have? How many options does he or she have? What is at stake for him or her? What does he or she expect?) and the various kinds of evidence available to him or her (Is the technology familiar? Have my past experiences with it been good? Does it look reliable? Does a known entity vouch for its reliability? Can I retaliate or complain if it does not work?) than to draw overly fine distinctions between types of objects of trust. It is also important for the designer to keep these questions in mind in case he or she wishes to create *un*certainty, *dis*trust, or doubt (the flip side of *Evidence for Trust*). The designer can help direct people's attention, bringing them to focus critically on some questions of reliance and not others. Anthropocentric and non-anthropocentric theories of trust both contribute to these practical questions about reliance and the reasons behind it. Although interpersonal trust is special, in these contexts it is also useful to consider a trust-like attitude that can be taken toward technologies and socio-technical systems.

Now that we have compared these views of what design for trust includes in its scope, we briefly consider Design for Values in relation to trust. Critical issues about the methodology of Design for Values and value-sensitive design are discussed elsewhere in this volume (see the chapters "▶ Value Sensitive Design: Applications, Adaptations, and Critiques," "▶ Design Methods in Design for Values," and "▶ Design for Values and the Definition, Specification, and Operationalization of Values"). Here we focus on two notes of critical caution specific to trust. First, it is important not to take too narrow a view of which potential trust relationships are relevant to a design. A residential community secured with forbidding gates and high walls may encourage trust among its residents, but discourage wider public trust among citizens. If a designer only looks at the effect on residents, they might deem this design to promote trust. However, a study with a wider perspective might judge that it hampers trust overall, partly because it enforces security physically instead of leaving it as a matter for the broader community to manage through a sense of mutual reliance and common purpose (a point that relates to the idea of *Discretion of*

*the Trusted* from earlier). Manders-Huits (2011) makes the important point that value-sensitive design does not always make it clear who is a stakeholder requiring consideration from the perspective of values. From an ethical point of view, all of those affected by a design are potentially relevant. This is also true regarding the value of trust: one should begin with a wide view of whose trust is relevant and what objects of trust are relevant.

Second, it is important not to separate the *process* of design for trust from its outcome. How one involves stakeholders, clients, and users in a process of Design for Values can have an effect on whether they form trust in and within the system when the design is actually implemented. The lessons of participatory design and stakeholder involvement are crucial (Reed 2008). Participatory processes can facilitate trust. For example, as one pair of authors writing about participatory design in architecture writes, "it may be only after clients understand what architects face in designing that trust develops. That is one advantage of a participatory design process ... where people have direct experience of the design challenges architects face" (Franck and Von Sommaruga Howard 2010). Here, it is the "experience" of Design for Values that reflexively stimulates trust in design.

## Cases and Examples

In this section we consider some concrete examples and case studies of design for the value of trust. First, it is useful to note these case studies do not explicitly mention or attempt to use the methodology of Design for Values described above (with the exception of Vermaas et al. 2010). A second general observation is that most of the examples and case studies come from the domain of ICT. Case studies in ICT often emphasize the importance of user identity mediation: how information about other users and their actions is mediated by the technology. For example, Pila (2009) and Carusi (2009) discuss how systems designed to distribute scientific knowledge and medical tasks can encourage (justified) user trust, focusing on the case studies of CalFlora (Van House 2002) and eDiaMoND (Jirotka et al. 2005). CalFlora is a library of botanical photographs and reports contributed by and available to scholars and horticulturalists. Van House raises the question of how the digital environment of CalFlora mediates judgments about the authenticity and authorship of the photos in the database, relating this to trust. She argues that an important condition for trust in such an environment is that contributors to the database have a stable virtual identity that can serve as a nominal pigeonhole in which to store knowledge about trustworthiness and authenticity (Van House 2002). Pila endorses this idea, but argues furthermore that having too much identifying information can cause users to rely on others in ways that emphasize existing biases, personal ties, and power relations, which distorts their judgment and blocks healthy skepticism (Pila 2009). The design of the system mediates these epistemic practices of trust and skepticism, in a way that links with the issue of *Evidence for Trust* mentioned earlier.

Pettit (2004) takes a gloomier view of trust online, arguing that the anonymity of those who interact online is a barrier to trust. Basing his argument on an account of

trust that links it strongly to the socially created and maintained currency of reputation (Pettit 1995), he argues that systems that do not allow for users to develop reputations make trust impossible. He argues that "on the Internet ... we all wear the ring of Gyges," referring to the tale in Plato's *Republic* in which a shepherd becomes invisible and takes advantage of his ability to engage in unjust acts with impunity (Pettit 2004, p. 118). For that reason, reputation and trust are both impossible on the Internet. Although Pettit's argument has been criticized (de Laat 2005), there is something highly valuable in it, as one can see from the fact that major commercial websites such as Amazon and eBay make reputational information storable and visible to buyers, and this has been a central feature of their sites over many years (see Resnick and Zeckhauser 2002). The system recreates this aspect of non-virtual trust within the virtual market environment, mediating the identities of users to one another.

Carusi (2009) also points out that it is important to build non-virtual aspects of trust into collaborative work-support systems. She focuses on the case study of eDiaMoND, a system for the distribution of breast cancer screening tasks among professionals. The eDiaMoND system allows judgments about medical images to be performed remotely as well as double-checked by professionals from another health-care facility. Jirotka et al. (2005) formulate the epistemological problems raised by the design of the system: "how can a reader who lacks knowledge of the (local) conditions of a mammogram's production read that mammogram confidently[?] ... second, how can a reader unknown to one be trusted to have read mammograms in an accountably acceptable manner?" (389, cited in Carusi 2009, p. 31). Carusi argues that when the system allows users to create and leave familiar kinds of contextual information, such as notes on a particular mammogram (even if they are anonymized), this helps professionals to situate their judgments of trust and distrust within familiar epistemic and social practices. New information technologies tend to disrupt these familiar practices and thereby bring issues of trust to the foreground. It may be necessary to build these practices into the system in some way in order to allow for trust among users.

User identity mediation and *Evidence for Trust* are major issues throughout these case studies and others. Describing an article by Bicchieri and Lev-On (2011) that looks at the effect of information about other agents on cooperative outcomes in a computer model of cooperative behavior, Vermaas et al. state that designers of ICT must confront questions of "which information about reputations, history, and identity should be made available ... how much ... and when" (2010, p. 498). Case studies from ICT thus clearly indicate that how a system constructs and mediates users' identities, motives, and reputation, and how much it allows them to provide trust cues to other users, is one of the primary issues of design for trust.

Some of the insights of these case studies of user identity mediation in ICT can be extended to architecture, urban planning, and other areas of design. For example, Katyal discusses how architecture can be used to encourage trust. "As architects bring natural surveillance to an area, they may ease community-police tensions" by encouraging mutual trust (2002, p. 1073). More generally, "architects can create spaces that bring people together or ones that set them apart. They can reinforce

feelings of familiarity and trust or emphasize harshness and social chaos" (2002, pp. 1086–1087). Like ICT, architecture mediates human relations, although it does so in physical space, where informational channels about other people are less rigidly controlled (and also less easy to change once the design has been brought into being).

User identity mediation is not the only issue concerning interpersonal trust in ICT, however. As mentioned in section "Comparison and Critical Evaluation" above, the trust of stakeholders who are not "users" in any standard sense is often highly relevant. For example, a system can facilitate the trust of an external party that has an interest in how the system functions. Vermaas et al. (2010) consider the design of ICT systems that allow companies to control and report commercial activities subject to tax and customs on behalf of the government tax authority. In this case, trust is relevant because the tax authority depends on the companies themselves to monitor and enforce the relevant laws and regulations on their transactions. The process of doing so is largely carried out by complex ICT systems. Vermaas et al. suggest that in these kinds of cases, what is needed is participatory involvement of the external control and regulatory body in the design of the system, ensuring that the system facilitates trustworthy behavior by companies who use it.[2] Other examples also indicate the importance of trust for nonusers: the design of an electronic voting machine system should encourage justified trust, not just among voters, but among the wider public, government bodies, etc. (see Pieters 2006).

So far we have focused on case studies of *interpersonal* trust as realized in and mediated by technology. There are other cases concerning trust *in technology*, concerning how people make willing choices to rely on complex technology under time pressure with limited evidence about reliability. This process is highly subject to influence by designers, because designers can use the technology and its embedding to communicate with users, establishing and/or building on normative expectations these users already have. Consider the examples from earlier in the paper of medical imaging technology used to help patients and physicians make diagnoses and decisions. The design of such systems mediates how patients and physicians rely on them. If a system is designed, as in Verbeek's (2008) example, so that an electronic image is seen and partially understood by the patient, then it provides a source of information independent from the physician's interpretation of the image. It can even be designed to deliver a written or symbolic message directly to the patient, to which the physician is a bystander. This creates complex relationship of reliance between patient, system, and physician. In addition to mediating patient-physician interpersonal trust, then, such systems also invite trust in technology itself. And despite anthropocentric accounts that emphasize

[2]Decentralized processes of control and regulation have developed greatly over the past forty years (Power 2007), and this has coincided with the development and integration of ICT systems in virtually all financial and business processes, so we can expect that similar kinds of cases, and similar issues of trust, will also appear in other regulatory and institutional contexts.

interpersonal trust, here the patient's reliance on the system has complex normative aspects. Ideas of the technology's purpose and function help determine what the patient expects of the system and what they rely on it to do. In the context of the practice of medicine and its associated ethical responsibilities, such a technology invites a normative, even moralized notion of trust.

## Open Issues and Future Work

There are several areas where more research is needed to advance the idea of design for the value of trust. Three areas will be highlighted here. First, our normative and moral expectations of other persons and entities are part of our reason for trusting them. How are such expectations about technology learned and communicated, and how do they guide our interaction with technology and other technology users? A framework is needed for thinking about the role of the designer as a communicator of expectations, affordances, and norms that guide us in *when* and *how* to rely on technology and when and how to rely on other technology users whose identities are mediated by the technology. This would involve interdisciplinary attention from philosophy, psychology, and design theory.

Second, returning to the issue of *Evidence for Trust*, what evidential standard should we try to meet in providing people with the materials for trust? What counts as the right amount and kind of evidence for people in a position to rely on technology and its other users? Since users are often under time and resource pressure, what simple cues can be used that also indicate or "ground" genuine trustworthiness? Some recent work suggests that the relevant standard is one on which the potential trustor should have an "adequate, sound justification for her trust" in a given technological system (Nickel 2013). But this may be too conservative a standard, because if the potential trustor has to wait for a sound justification, it may actually inhibit his or her trust formation. To what extent should we allow or even encourage people to make a leap of faith, counting on the reliability of others to catch them? If we do encourage people to make that leap, then we should be sure that the technology really lives up to their (reasonable) expectations of reliability.

A third area of future research concerns technological artifacts that involve built-in social and linguistic attributes that invite interpersonal trust (e.g., a talking robot with a friendly face). To what extent may we design anthropomorphic trust-inviting attributes such as these for users who cannot discern that they are not "real," e.g., children or severely mentally disabled persons? How does such technology need to be framed and implemented in order to be respectful to technology users? For example, who is responsible for what the robot says? Should we make such technology directly *responsive* to user expectations (e.g., about what values such as sustainability are implemented in the interface of a car dashboard) in order to enhance trust?

## Conclusions

This chapter has explored design for trust, focusing on how our conceptualization of trust affects what we take "design for trust" to mean. If we take interpersonal trust as the only kind of trust, then technology's role is to mediate interpersonal trust. The area of application in which this is most apparent in case studies is user identity mediation, the way in which users of a technological system are presented to other users. However, design can also mediate human interpersonal trust in other ways. For example, it can do so by changing relationships (e.g., between patient and physician or between a private company and a government agency) or by establishing a new relationship (e.g., between a homeowner in a gated community and a stranger from outside that community). Furthermore, it is useful to consider designed artifacts and systems as being the object of trust. This is accentuated even further as technology and design are used increasingly to mediate central elements of our lives such as friendships and family relationships, work, mobility, and political participation. Future empirical and theoretical work is needed to understand better what we owe to those who rely on design in order to foster and to provide *sound*, *well-grounded* support for trust in technology.

## Cross-References


- ▶ Design for the Value of Presence
- ▶ Design for Values in ICT
- ▶ Design Methods in Design for Values
- ▶ Design for Values and the Definition, Specification, and Operationalization of Values
- ▶ Value Sensitive Design: Applications, Adaptations, and Critiques


## References


Akao Y (ed) (1990) Quality function deployment: integrating customer requirements into product design. Productivity, Cambridge

Baier A (1986) Trust and antitrust. Ethics 96:231–260

Ben-Ner A, Halldorsson F (2010) Trusting and trustworthiness: what are they, how to measure them, and what affects them. J Econ Psychol 31:64–79

Bicchieri C, Lev-On A (2011) Studying the ethical implications of e-trust in the lab. Ethics Inf Technol 13:5–15

Carusi A (2009) Implicit trust in the space of reasons and implications for technology design: a response to Justine Pila. Soc Epistemol 23:25–43

Coleman J (1990) Foundations of social theory. Harvard University Press, Cambridge, MA

Cross N (2006) Designerly ways of knowing. Springer, London

Cummings ML (2006) Integrating ethics in design through the value-sensitive design approach. Sci Eng Ethics 12:701–715

de Laat PB (2005) Trusting virtual trust. Ethics Inf Technol 7:167–180
Eyal N (2012) Using informed consent to save trust. J Med Ethics. doi:10.1136/medethics-2012-100490
Franck KA, Von Sommaruga Howard T (2010) Design through dialogue: a guide for architects and clients. Wiley, Chichester
Friedman B (1995) "It's the computer's fault" – reasoning about computers as moral agents. In: Conf Companion of CHI 1995, ACM Press, pp 226–227
Friedman B, Kahn PH Jr, Howe DC (2000) Trust online. Commun ACM 43:34–40
Friedman B, Kahn PH Jr, Borning A (2006) Value sensitive design and information systems. In: Zhang P, Galletta D (eds) Human-computer interaction and management information systems. M.E. Sharp, New York, pp 348–372
Gambetta D (1988) Can we trust? In: Gambetta D (ed) Trust: making and breaking cooperative relations. Basil Blackwell, Oxford, pp 213–237
Glass A, McGuinness DL, Wolverton M (2008) Toward establishing trust in adaptive agents. In: Proceedings of the Conference on Intelligent User Interfaces (IUI), pp 227–236
Hardin R (1993) The street level epistemology of trust. Polit Soc 21:505–529
Hardin R (2006) Trust. Polity, New York
Jirotka M, Procter R, Hartswood M, Slack R, Simpson A, Coopmans C, Hinds C, Voss A (2005) Collaboration and trust in healthcare innovation. The e-DiaMoND case study. Comput Support Collab Work 14:369–398
Jones K (2004) Trust and terror. In: DesAutels P, Urban Walker M (eds) Moral psychology: feminist ethics and social theory. Rowman and Littlefield, Lanham, pp 3–18
Katyal NK (2002) Architecture as crime control. Yale Law J 111:1039–1139
Kelton K, Fleischmann KR, Wallace WA (2008) Trust in digital information. J Am Soc Inf Sci Technol 59(3):363–374
King B (1989) Better design in half the time: implementing QFD in America, 3rd Ed. GOAL/QPC, Methuen
Kiran AH, Verbeek P-P (2010) Trusting our selves to technology. Knowl Technol Policy 23:409–427
Lagerspetz O, Hertzberg L (2013) Trust in Wittgenstein. In: Mäkelä P, Townley C (eds) Trust: analytic and applied perspectives. Rodopi, Amsterdam, pp 31–51
Lee G, Lee WJ (2009) Psychological reactance to online recommendation services. Inf Manag 46(8):448–452
Manders-Huits N (2011) What values in design? The challenge of incorporating moral values into design. Sci Eng Ethics 17(2):271–287
McLeod C (2002) Self-trust and reproductive autonomy. MIT Press, Cambridge, MA
Möllering G (2006) Trust: reason, routine, reflexivity. Elsevier, Amsterdam
Nickel PJ (2010) Horror and the idea of everyday life: on skeptical threats in *Psycho* and *The Birds*. In: Fahy T (ed) The philosophy of horror: philosophical and cultural interpretations of the genre. University of Kentucky Press, Louisville, pp 14–32
Nickel PJ (2011) Ethics in e-trust and e-trustworthiness: the case of direct computer-patient interfaces. Ethics Inf Technol 13:355–363
Nickel PJ (2013) Trust in technological systems. In: de Vries MJ, Hansson SO, Meijers AWM (eds) Philosophy of engineering and technology, vol 9, Norms in technology., pp 223–237
Nickel PJ, Franssen M, Kroes P (2010) Can we make sense of the notion of trustworthy technology? Knowl Technol Policy 23:429–444
O'Neill O (2002) Autonomy and trust in bioethics. Cambridge University Press, Cambridge
Pettit P (1995) The cunning of trust. Philos Public Aff 24:202–225
Pettit P (2004) Trust, reliance and the internet. Anal Krit 26:108–121
Pieters W (2006) Acceptance of voting technology: between confidence and trust. In: Stølen K et al. (eds) Trust management. Lecture notes in computer science, vol 3986. Springer, Berlin, pp 283–297

Pila J (2009) Authorship and e-science: balancing epistemological trust and skepticism in the digital environment. Soc Epistemol 23:1–24
Power M (2007) Organized Uncertainty. Oxford University Press, Oxford
Reed MS (2008) Stakeholder participation for environmental management: a literature review. Biol Conserv 141:2417–2431
Resnick P, Zeckhauser R (2002) Trust among strangers in internet transactions: empirical analysis of eBay's reputation system. In: Baye MR (ed) The economics of the internet and e-commerce, Advances in applied microeconomics, vol 11.. Elsevier, Amsterdam, pp 127–157
Riegelsberger J, Sasse MA, McCarthy JD (2003) Shiny happy people building trust? Photos on e-commerce websites and consumer trust. CHI Lett 5:121–128
Roubroeks M, Ham J, Midden C (2011) When artificial social agents try to persuade people: the role of social agency on the occurrence of psychological reactance. Int J Soc Robot 3(2):155–165
Simpson E (2011) Reasonable trust. Eur J Philos. doi:10.1111/j.1468-0378.2011.00453.x
Smolkin D (2008) Puzzles about trust. South J Philos 46:431–449
Uslaner E (2002) The moral foundations of trust. Cambridge University Press, Cambridge
Van House N (2002) The CalFlora study and practices of trust: networked biodiversity information. Soc Epistemol 16:99–114
Verbeek P-P (2008) Obstetric ultrasound and the technological mediation of morality: a postphenomenological analysis. Hum Stud 31:11–26
Vermaas PE, Tan Y-H, van den Hoven J, Burgemeestre B, Hulstijn J (2010) Designing for trust: a case of value-sensitive design. Knowl Technol Policy 23:491–505

# Part IV

# Domains

# Design for Values in Agricultural Biotechnology

Henk van den Belt

## Contents



**Abstract**

Agricultural biotechnology dates from the last two decades of the twentieth century. It involves the creation of plants and animals with new useful traits by inserting one or more genes taken from other species. New legal possibilities for patenting transgenic organisms and isolated genes have been provided to promote the development of this new technology. The applications of biotechnology raise a whole range of value issues, like consumer and farmer autonomy, respect for intellectual property, environmental sustainability, food security, social justice, and economic growth. Hitherto the field has not yet witnessed any deliberate attempt at value-sensitive design or design for values. The reason is that under the influence of strong commercial motivations, applications have been developed first and foremost with simple agronomic aims in view, such as herbicide tolerance and insect resistance, traits which are based on single genes. The opportunities for value-sensitive design appear to be constrained by the special character of the biological domain. Many desirable traits like drought

H. van den Belt (✉)
Wageningen University, Wageningen, The Netherlands
e-mail: henk.vandenbelt@wur.nl; lizandhenk@gmail.com



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_23

tolerance are genetically complex traits that cannot be built into organisms by the insertion of one or a few genes. Another problem is that nature tends to fight back, so that insects become immune to insect-resistant crops and weeds become invulnerable to herbicides. This leads to the phenomenon of perishable knowledge, which also calls the so-called patent bargain into question. The possibilities for value-sensitive design will likely increase with synthetic biology, a more advanced form of biotechnology that aims at making biology (more) "easy to engineer." Practitioners of this new field are acutely aware of the need to proceed in a socially responsible way so as to ensure sufficient societal support. Yet synthetic biologists are currently also engaged in a fundamental debate on whether they will ultimately succeed in tackling biological complexity.

**Keywords**

Intellectual property • Complex traits • Sustainability • Trade-offs • Perishable knowledge • Synthetic biology

## Introduction

Modern agricultural biotechnology dates from the final decades of the last century. The adjective "modern" is sometimes added as an essential specification to distinguish contemporary biotechnology from age-old forms of human intervention with living nature such as traditional agriculture, conventional plant and animal breeding, and ancient fermentation techniques employed in making bread, beer, wine, cheese, and soy products. This linguistic usage may be slightly pedantic, however, as the lay public usually identifies "biotechnology" exclusively with its modern incarnation. While (modern) agricultural biotechnology is based on techniques of genetic engineering and thus involves manipulation on the level of DNA molecules, conventional breeding operates on the macroscopic ("phenotypic") level by selecting and crossing suitable individual organisms in order to create new varieties of plants and animals. As Charles Darwin already pointed out, agricultural practices automatically entail selection even when farmers do not consciously engage in deliberate breeding (Darwin 1972 [1859], p. 93). Similarly, while fermentation techniques are based on the activity of microorganisms (like yeasts), it is only since the investigations of Louis Pasteur that we are aware of this fact and that we can use our microbiological knowledge to improve these techniques. More recently, such knowledge has been supplemented with insights from genetics and molecular biology, thus opening a wide field of application to genetic engineering. Ancient fermentation techniques have thus gradually evolved into what is often called industrial biotechnology (or "white" biotechnology), which deals with the deployment of genetically engineered microorganisms and tailored enzymes for the optimization of industrial fermentation processes. In this chapter, I will however confine myself to agricultural biotechnology (sometimes referred to as "green" biotechnology). This area of biotechnology happens to be very controversial and to

raise many ethical concerns. It may therefore be worthwhile to explore the possibilities of value-sensitive design (VSD) and design for values in this particular field.

## Main Technologies

The so-called recombinant-DNA (or r-DNA) technology forms the technical core of modern biotechnology. It comprises a set of procedures by which segments of DNA from any organism can be "cut" at specific places and "pasted" together to form new recombinant-DNA molecules, which can be "inserted" into a recipient organism by using one or another method of gene transfer. This core technology uses two different groups of enzymes, which are involved in either "cutting" (restriction enzymes) or "pasting" (ligases). Those enzymes serve as the molecular scissors and glue by which genetic engineers perform their cut-and-paste work. There are a number of ways to insert the new r-DNA molecule (comprising one or more foreign genes from a distantly related or virtually unrelated organism) into the recipient organism. In the early years of genetic engineering, a relatively small piece of foreign genetic material (e.g., the human DNA sequence coding for the production of insulin) was often recombined with plasmids (i.e., circular pieces of bacterial DNA), and those plasmids were themselves used as vectors to be introduced into bacteria, where they would be replicated along with their bacterial hosts (bacterial cloning). The disadvantage of this method was that only relatively small stretches of DNA could be incorporated in plasmids. In the 1980s the introduction of the use of yeast chromosomes instead of plasmids allowed the multiplication or "cloning" of much longer segments of DNA. (Another breakthrough of the 1980s, the polymerase chain reaction or PCR, made it possible for the first time to multiply DNA sequences in vitro, that is, without the need to put those sequences into bacterial or yeast cells.) Other methods of gene transfer are microinjection, whereby the genetic material is injected into the host cell by means of a miniscule glass syringe, and bioballistics, whereby the foreign DNA is coated on tiny metal particles and then shot into the host cell with the aid of a device called a gene gun. Viruses and bacteria are also used as vehicles for transferring genes. In plant biotechnology, *Agrobacterium tumefaciens* has become a popular vector for mediating the transfer of genes to various plants. This bacterial species is a "natural genetic engineer" that transmits part of its own DNA to the plants it infects, all the while causing tumor crown galls at the wound sites. Modern plant biotechnology has turned disabled versions of this bacterium into a Trojan horse for the transfer of recombinant genes.

The insertion of a foreign gene into a particular organism only makes sense, of course, if the gene and its function or the protein for which it codes are known. The background knowledge about genes and their functions across a wide array of biological species and taxa is still expanding. Thanks to spectacular advances in the techniques for sequencing DNA, the human genome and the genomes of various other organisms have been completely mapped. The advance of genomics has also shown that many of the older theoretical conceptions of molecular biology (like the

Central Dogma that genetic information always flows from DNA via RNA to protein or the idea that a gene is always represented by a fixed stretch of DNA and always codes for one and only one type of protein) are often far too simplistic. In fact, the functioning of genomes has turned out to be exceedingly complex (Griffiths and Stotz 2013). The new discipline of bioinformatics has been brought into being to handle and process the enormous and constantly increasing mass of genomic data. Some of the new insights gained have resulted in new molecular tools, which can be deployed for the benefit of genetic engineering but also to enhance conventional breeding, as happens, for example, in marker-assisted selection (MAS).

## History

When the unprecedented possibilities that would be opened up by recombinant-DNA technology were first realized in the early 1970s, molecular biologists declared a temporary moratorium on further experiments in this area to discuss the possible consequences of this new line of work and to devise measures that would allow it to be continued in a responsible way. This unleashed a broad societal debate in which issues of safety and security but also more remote ecological and social consequences of r-DNA technology were extensively discussed. The researchers who were most directly concerned with this type of work finally agreed among themselves that r-DNA experiments could be conducted safely in the confined environment of specially secured laboratories, but large segments of the lay public remained unconvinced. At the end of the 1970s, however, biotechnology was increasingly perceived by governments as an exciting field of technological innovation that would lead to renewed economic growth and restore international competitiveness for western countries. The production of human insulin by genetically modified bacteria, realized in 1978 by the first biotech company, Genentech, was the key event that aroused high expectations.

The huge economic potential of this new field of technology would however only be unlocked, it was thought, if biotechnological inventions were to receive proper legal protection. In the landmark case of *Diamond v. Chakrabarty* concerning the patentability of a genetically modified oil-consuming bacterium, the US Supreme Court ruled in 1980 that "anything new under the sun that is made by man," whether living or nonliving, can be patented. In subsequent years US jurisprudence explicitly extended patentability to multicellular organisms like plants (1985), oysters (1987), and mammals (1988). Other western countries ultimately followed the American example, albeit with some delays and hesitations. In 1988 the patent offices of the USA, the European Union, and Japan proclaimed the new policy line that DNA sequences and genes would also be eligible for product patents. Their justification was that sequences and genes, when isolated and purified, would be essentially different from their natural counterparts and therefore qualify as inventions rather than discoveries. (This standpoint was later incorporated in the European Directive on the Legal Protection of Biotechnological

Inventions of 1998, Directive 98/44/EC.) In a parallel move, the legal protection of plant varieties resulting from conventional breeding by so-called plant breeders' rights would also be tightened up. In 1961 a handful of western (mainly European) countries had concluded the first international agreement on plant variety protection, called UPOV after its French acronym (*Union internationale pour la Protection des Obtentions Végétales*). This agreement gave the originators exclusive rights on commercializing their plant varieties but granted other breeders the right to use these varieties as starting material for further breeding (breeder's exemption) and left farmers the freedom to save seed from their harvest for the next planting season (farmer's privilege). In 1991 a new international agreement was concluded (referred to as UPOV 1991), which drastically curtailed the breeder's exemption and virtually annulled the farmer's privilege, bringing plant breeders' rights more in line with patent law (GRAIN 2007). In the eyes of its main beneficiaries, the intellectual property regime also needed to be globalized. Driven by an influential business lobby in the pharmaceutical, biotech, and entertainment industries, the US and European governments used their clout in international trade negotiations to "persuade" reluctant developing countries to accept the (for them often disadvantageous) terms of the *TRIPS Agreement*, which was concluded in 1994 as part of an overall WTO package. The TRIPS agreement (standing for *Trade-Related aspects of Intellectual Property rights*) sets worldwide minimum standards for the protection of intellectual property rights (including patents, copyright, and breeder's rights). It mandates that, with few exceptions, "patents shall be available for any inventions, whether products or processes, in all fields of technology" (art. 27.1). Countries are allowed to exclude plants and animals (other than microorganisms) from patentability, but "Members shall provide for the protection of plant varieties either by patents or by an effective *sui generis* system or by any combination thereof" (27.3b). Breeder's rights are an example of a *sui generis* system of plant variety protection. Many developing countries have meanwhile joined the UPOV 1991 agreement to fulfill their TRIPS obligations. In the USA and the European Union, genetically modified crops may even be doubly protected by patents and by plant breeder's rights.

Legislation on intellectual property rights is only part of the legal framework regulating biotechnology. The recombinant-DNA controversy of the 1970s had been "resolved" (or at least temporarily closed) by the introduction of strict safety and security rules for the research labs in which gene-splicing experiments were to be conducted. This set of rules obviously no longer sufficed when in a next stage the new technology was also applied to the creation of transgenic crops and farm animals, which had to grow and live in much less confined settings than secured labs and were often ultimately destined to enter the human food chain. The first field trials with GMOs (genetically modified organisms) occurred in the 1980s. This new stage in the development of biotechnology presented a huge challenge to the regulatory authorities. Largely for historically contingent reasons, the United States and the European Union have devised sharply contrasting policy answers to this challenge. The US response, by and large, has been to treat agricultural and food biotechnology as business as usual, to regulate its final products in the same manner

as those of any other technology, and to declare the process by which the products are made (i.e., genetic engineering) irrelevant for regulatory purposes. This type of policy response has been characterized as the "product frame" (Jasanoff 2005). The EU approach to regulation has been completely different. European policymakers consider agricultural and food biotechnology as more than just business as usual and see the process of genetic modification itself as a relevant factor for regulation. This type of policy response has been characterized as the "process frame" (Jasanoff 2005). It implies a much more "precautionary" approach to the possible ecological risks and health hazards of GMOs, separation of GM and non-GM (conventional and organic) product flows, monitoring and traceability, and mandatory labeling of GM foods to ensure freedom of choice to consumers. A similar approach was adopted by policymakers in Japan and South Korea. On the global scale, the confrontation of these two opposed policy frames has led two "regulatory polarization" (Bernauer 2003) and given rise to fierce disputes before the World Trade Organization.

Transgenic crop varieties were first commercialized in 1996. Since then a suite of different GM crops have spread to different parts of the world in a rather uneven pattern, determined by varying socioeconomic and agroecological conditions but also by different regulatory frameworks and intellectual property arrangements. The area planted with biotech crops has increased 100-fold from 1.7 million hectares in 1996 to 170.3 million hectares in 2012 (James 2012). The two traits that have most often been inserted into GM varieties are herbicide tolerance and insect resistance. The main agricultural crops involved are soybean, canola, maize, and cotton. Transgenic crops are mostly grown in North and South America and in Asia (especially China and India), while Europe and Africa are the continents with a very low adoption rate in terms of the number of approved varieties as well as of planted area. In Europe, stringent regulation and public distrust are retarding factors, while in Africa it is the very lack of indigenous regulatory capacity and the fear of losing product markets in Europe, along with a shortage of GM crops suitably adapted to African agroecological conditions, which explain the low adoption rate. Adoption may also be influenced by the vicissitudes of intellectual property protection and biosafety regulation, as is illustrated by the case of GM soybeans in South America. At an early stage, Argentina eagerly adopted the so-called "Roundup-Ready" soybean, which had been developed by the US company Monsanto as a GM variety resistant to its proprietary herbicide glyphosate (trade name "Roundup"). The variety was actually without legal protection in Argentina and therefore formally in the public domain, as Argentine law did not allow patents on plants and Monsanto had failed to apply for a plant breeder's right (Correa 2006). This did not prevent Monsanto to claim royalties from Argentina for the use of its "proprietary technology." The US company even went so far as to seize shiploads of Argentine soy meal in European ports and sue for patent infringement there (in the end, European courts rejected Monsanto's claims). Through illegal smuggling from Argentina, glyphosate-resistant soybeans also reached farmers in Paraguay and Brazil, where the new GM variety had not yet been approved by the regulatory authorities. Widespread adoption by farmers in

those countries created a *fait accompli*, which was subsequently legalized by a formal approval not based on a careful biosafety assessment. Something similar happened in India with insect-resistant Bt cotton (containing a gene from *Bacillus thuringiensis* that produces a toxin against insects). This variety had been developed by Monsanto and its Indian subsidiary Mahyco. These companies proved unable to retain their intellectual property control over the new variety, after Gujarat farmers had somehow appropriated the transgenic seeds (possibly from testing fields), crossed it out with indigenous varieties, and in the process created a huge market for "stealth seeds" (Herring 2007). The farmers' actions also defied the Indian supervisory agency charged with biosafety regulation, but state and federal authorities in India did not dare to alienate their farmer constituencies by ordering the destruction of GM cotton harvests.

## Values and Value Issues

Agricultural biotechnology is still highly controversial. Some objections are ostensibly based on religious views, like the charge that by crossing species boundaries, man is *playing God*. There is also the stigma of *unnaturalness* that is often attached to GMOs. Both charges can be readily combined, as is shown in the work of Jean-Jacques Rousseau, the precursor of modern romanticism. His *Émile ou de l'é ducation* can actually be read retrospectively as a radical condemnation of modern biotechnology. The opening sentence states: "God makes all things good; man meddles with them and they become evil" (Rousseau 1966 [1762], p. 35). In a later chapter, Rousseau claims that it is definitely nature's (or God's?) intention to keep the various species apart and distinct: "The insurmountable barriers that Nature has placed among the various species, so that they will not become mingled, make her intentions abundantly clear. She has not simply established order: she has also taken effective measures to prevent it from being disturbed" (ibid., p. 359). There is thus no doubt that Rousseau would have opposed attempts to recombine genetic material from different species.

Echoes of Rousseau's romantic celebration of nature are ubiquitous today, especially in food adverts (Doorman 2012). Thus a big Dutch dairy company advertises one of its products, "pure" milk from cows in grassy pastures, as "milk such as Nature intended it to be." But the proponents of modern biotechnology can also speak Rousseau's language, as is testified by the CEO of the industrial biotech company DSM, Feike Sijbesma, in an interview on second-generation biofuels: "We are on the threshold of a green revolution to return to a society living from and with nature" (quoted in Banning 2012). The fact that the most divergent causes can apparently be justified by an appeal to nature should make us wary of the validity of any argument invoking the presumed naturalness or unnaturalness of GMOs.

A value that enjoys wide endorsement in liberal-democratic market economies is *freedom of choice* for consumers. Even if one does not share the religious objections against biotechnology or finds the charge of unnaturalness inappropriate, one would still grant people the right to act on their personal convictions in their personal lives.

Mandatory labeling of GM foods might be seen as a straightforward way to secure this right. If consumers are to be given a free and informed choice between GM and non-GM foods, however, then in practice a compromise has to be struck. It is impossible to guarantee that foods that are not labeled "GM" will be 100 % GM-free in situations where both categories of food are admitted to the market. The solution is to set an upper threshold for "contamination" (such as the 0.9 % limit in the EU). A further departure from an ideally free choice for the consumer results from the fact that the whole regulatory machinery set up to secure this choice – proper distances between fields with GM and conventional or organic crops ("coexistence"), separation and tracing of product flows, and adequate liability rules – tends to discourage the production and marketing of GM foods. Ultimately the consumer may end up having only the option of "choosing" non-GM foods. At present, this is by and large the situation in the EU. From a moral point of view, it is far from ideal. In the USA, the situation with regard to consumer freedom is entirely different but not ethically more satisfactory. Here the adopted regulatory model (the "product frame") has effectively excluded mandatory labeling and thus denied consumers the possibility to exercise their right of an informed choice between GM and conventional foods.

The EU regulatory framework is predicated on the assumption that "coexistence" and separation are viable options. Critics of agricultural biotech contest that assumption. Wherever GM crops are being grown and processed into food or other products, inadvertent transfer of transgenes to conventional crops and weedy relatives through pollen transport and the mixing up of seeds in the processing chain are bound to occur. Organic farmers in particular can be economically harmed when their harvest becomes "contaminated" and no longer satisfies customary certification requirements. The risk of contamination becomes especially troublesome with some newer generations of biotech crops. It would hardly be acceptable, for example, when GM plants engineered to produce "biopharmaceuticals" entered the human food chain. Researchers and biotech companies are therefore exploring various possibilities of biological containment (e.g., by making GM seeds sterile) to address the issue of unwanted fallout from the growing of GM crops. The old ethical principle of Hippocrates may apply here: *First, do no harm!*

A value to which biotech companies attach great importance is *respect for intellectual property*. For them, patents and plant breeder's rights are a just reward for their inventive efforts and allow them to recoup the costs and expenses incurred in creating new GM varieties. Hence they very much lament any unauthorized use of "their" technologies, for example, by farmers who grow "pirated" GM crops without paying them any royalties. Although patents, plant breeder's rights and other intellectual property rights are territorially based, it is striking that companies tend to see their inventions as proprietary also in those countries in which no patents or breeder's rights have been filed. Thus Monsanto claims royalties on the use of GM soybeans in Argentina even though their invention is not legally protected in that country. It is also not unusual for biotech companies to magnanimously "donate" their technologies to humanitarian initiatives for use in countries where they have no markets (as with the WIPO World Intellectual Property Organization

Global Responsibility Licensing Initiative), but what exactly do they give away if they have no patents in such countries in the first place? For farmers, property rights are also at stake, but their concern is rather that modern *intellectual* property threatens to erode their *tangible* property. In the old days, when a farmer bought seed from the seed merchant, it truly became his property, that is, he could do with it whatever he liked. He could use it to grow his crop and save seed from the harvest for replanting in the next season (or he could exchange it with his neighbor or even sell it on the market). This age-old practice of seed saving (and seed exchange) has traditionally been at the core of an informal system of crop improvement (De Schutter 2011). Even the first international agreement on plant breeder's rights (UPOV 1961) still recognized the farmer's privilege or the right to save seed on-farm. The rise of agricultural biotechnology would drastically change this. New interpretations of patent law, followed by a drastic revision of plant breeder's rights (UPOV 1991), no longer allow on-farm seed saving. When a farmer buys GM seed from a biotech seed company, it no longer becomes his full property because he no longer acquires the right to make use of an inherent biological characteristic of the seed, i.e., its natural capacity to reproduce itself. In fact, it would be more appropriate to say that the farmer "rents" the GM technology incorporated in the seed for the duration of only one growing season. Or as was stated in a US Supreme Court case, the biotech company "sells the seeds subject to a licensing agreement that permits farmers to plant the purchased seed in one, and only one, growing season" (Bowman v. Monsanto Co. 2013).

While biotech companies demand respect for intellectual property, others fear that the *autonomy* and *independence* of farmers will be increasingly undermined by more stringent IP restrictions on saving seed. The famous report on the International Assessment of Agricultural Knowledge, Science and Technology for Development (IAASTD) expresses "concern about present IPR instruments eventually inhibiting seed-savings and exchanges" (IAASTD 2008, p. 42), thereby restricting the capability of farmer communities to develop locally adapted varieties and to maintain gene pools through in situ conservation – essential to local practices that enhance food security and sustainability (ibid., pp. 43–44).

Debates on agricultural biotechnology also turn on values like *environmental sustainability*, *food security*, *energy security*, *social justice*, *health*, *wealth*, and *economic growth*. The beauty of this beneficent technology, according to some of its adherents, is precisely that it allows us to have it all. From the very outset, biotech champions have raised expectations about unlimited wealth creation along with promises about incredibly benign environmental and socioeconomic effects. It is not unusual, of course, for newly emerging technologies to fuel high expectations, but in the case of agricultural biotechnology, the "cycles of hype and hope" seem to be exceptionally tenacious. In 2008 Hugh Grant, CEO of Monsanto, stated that in his view sustainability means that "we produce more and conserve more simultaneously" (Grant 2008). Biotech makes this possible. It allegedly allows us to produce more food, more feed, more fiber, and more energy all at once and also to protect the environment, thus finally enabling us to escape from Hermann Goering's eternal dilemma of guns versus butter. No hard choices are necessary.

Rather than the grim choice "food *or* fuel," so much impressed upon us by the backlash caused by first-generation biofuels, we can have "food *and* fuel" (and much else besides). Even more, agricultural biotechnology contains an internal code that is inherently pro-poor: "The novel thing about biotech is that it's scale neutral. Seeds deliver scale neutrality whether you're a one-acre smallholder in Uganda or a 1,000-acre grower in the Mississippi Delta. And the benefits of using biotech seeds are roughly the same" (ibid.). That may sound too good to be true, and it probably is (for a criticism of the hidden assumptions behind the framing of agricultural biotechnology as pro-poor, see Scoones (2002) and Glover (2009)). At any rate, in 2008 Grant held out the prospect of a new generation of *drought-tolerant* maize varieties, which his company intended to launch in the US Midwest in 2012 or 2013, and which it would subsequently make available to farmers in sub-Saharan Africa with the least possible delay through the WEMA (Water Efficient Maize for Africa) public-private partnership. Meanwhile, African regulatory capacity is being built up in the form of the African Biosafety Network of Expertise (ABNE) in Burkina Faso, nominally an "Africa-based, Africa-owned, and Africa-led" initiative (Vaidyanathan 2010), but funded by the Bill and Melinda Gates Foundation. This network is supposed to smooth the way for the arrival of drought-tolerant GM crops.

In December 2011, the US Department of Agriculture "deregulated" (approved) Monsanto's so-called *DroughtGard* maize, a GM maize variety containing the cold-shock protein gene *cspB* derived from the bacterium *Bacillus subtilis*, which is said to confer drought tolerance. WEMA expects to release adapted versions of this transgenic drought-tolerant maize in sub-Saharan Africa as early as 2017 (James 2012, p. 10). Given the increasing vulnerability of agricultural harvests to extreme weather conditions due to climate change, enhanced drought tolerance of crops might be considered a highly desirable trait. The same holds for improved water use efficiency (WUE), in view of the circumstance that already 70 % of global freshwater is currently being used by agriculture. The key question is whether and to what extent agricultural biotechnology can indeed contribute to the alleviation of periodic drought stress and water scarcity.

## Hype Versus Caution

There are several reasons for striking a skeptical or at least cautionary note with regard to the expected environmental and socioeconomic performance of new generations of GM crops. It is a sobering thought that the possibility to insert genes for nitrogen fixation derived from nitrogen-fixing bacteria into nonleguminous crops was already announced in 1981 as a "promise" of the new biotechnology, that this possibility has not been realized until now, and that a recent forecasting exercise (Charles et al. 2010) sets the expected arrival of nitrogen-fixing GM crops beyond a 20-year time interval. So when this early promise is finally realized (if it is to be realized), it will have taken more than 50 years! Surely it would be extremely attractive, both from a socioeconomic and environmental point

of view, to have the trait of nitrogen fixation in our crops. Other early promises already made in 1981 were drought tolerance and salt tolerance of plants. But perhaps biotech companies had other priorities during the past 20 or 30 years, such as making crops resistant to their proprietary herbicides – as Monsanto first did by creating GM "Roundup-Ready" varieties of canola, maize, and soybean that would tolerate its registered glyphosate herbicide, an example of strategic behavior that was to be quickly followed by its main competitors Syngenta, DuPont, Bayer, and BASF and that clearly made economic sense (Harhoff et al. 2001).

It is also plausible that biotechnology is as yet simply unable to deal with serious biological complexity. We have to take account of the fact that the two traits that have been introduced into the currently most widely used GM crops – herbicide tolerance and insect resistance – are relatively simple *single-gene* traits. Traits such as drought tolerance, salt tolerance, and other forms of abiotic stress tolerance (heat tolerance, cold tolerance, light tolerance, etcetera), by contrast, are (genetically and physiologically) *complex* traits involving many genes and complex gene-environment interactions. Moreover, there is also a subtle interplay between different abiotic stress conditions, occasionally reinforcing or mitigating each other. As a recent review article summarized, "The acclimation of plants to abiotic stress conditions is a complex and coordinated response involving hundreds of genes. These responses are also affected by interactions between different environmental factors and the developmental stage of the plant . . ." (Mittler and Blumwald 2010, p. 444). There may therefore be reasonable doubt about the claim that agricultural biotechnology can come to grips with this complexity, despite Hugh Grant's confident announcement that Monsanto's drought-tolerant maize varieties will alleviate production losses from periodic drought occurring in the American Midwest (Grant 2008). Grant referred to field trials showing yield increases of 8–10 % "in dry land corn [maize] environments," but this is of course no guarantee that the same yield increases will actually be obtained in the maize fields of Midwestern farmers, and still less so in sub-Saharan maize fields (African Centre for Biosafety 2013). Supporters of transgene-based drought tolerance have already adjusted their initial high expectations downward (Edmeades 2013, p. 27). It is not even sure that GM drought-tolerant crops will ultimately turn out to be the best answer to the problem of drought stress. Significantly enough, other companies including Monsanto's biotech rivals DuPont and Syngenta have meanwhile launched drought-tolerant maize varieties, in which the desired trait has not been created by genetic engineering but by conventional breeding informed by the molecular technique of marker-assisted selection (Edmeades 2013, pp. 16–20). A critical report issued by the Union of Concerned Scientists concludes that transgene-based drought-tolerant maize is not superior to maize in which this trait has been obtained through conventional means, that Monsanto's variety offers modest protection only under moderate but not under severe drought conditions, and that it shows no advantages at all with regard to water use efficiency (Gurian-Sherman 2012). The lackluster performance of these new biotech maize varieties should not be surprising, as drought tolerance is controlled by many different genes and genetic engineering so far has manipulated only a few genes at a time.

There is a further reason to be cautious about claimed and expected environmental benefits of GM crops. This reason may be summed up in the slogan: *Nature fights back* (Carson 1962, Chap. 15). While environmentalists' fears about GM crops often concentrate on the risk that transgenes outcross with wild plants and inadvertently create nasty superweeds, there is also the classical Darwinian scenario that continued use of certain herbicides on a massive scale, enabled and even encouraged by the herbicide tolerance engineered into the crop plants themselves, could act as a selection pressure favoring the development and spread of resistant weeds. What is currently happening in US soybean and maize cultivation is a case in point. Monsanto's "Roundup-Ready" (glyphosate-tolerant) crops have been immensely successful in the USA, where they currently cover 90 % of the soybean area and 80 % of the maize area. In comparison with some older and more aggressive herbicides, glyphosate is relatively benign in its effects on wildlife. Another environmental advantage is that the combined use of glyphosate and glyphosate-tolerant crops enables many farmers to practice low-tillage agriculture, with much less soil degradation and fuel use. Many successive years of glyphosate use, however, have now resulted in at least nine nasty weed species that have gained immunity to this herbicide. The expectation is that by 2015 some 40 % of the cultivation area will harbor resistant weeds. Farmers have to resort to older and less ecologically benign herbicides such as 2,4-D and dicamba, in addition to using Roundup, to kill the new invaders. Agrochemical and biotech companies are meanwhile developing new herbicide-tolerant varieties of soybean and maize with "stacked" transgenes that will not only tolerate glyphosate but also other herbicides (Kilman 2010; Keim 2012). We are thus witnessing an ongoing "arms race" between biotech and nature, which shows that the environmental benefits of agricultural biotechnology are sometimes only temporary rather than durable or truly "sustainable."

For the biotech companies involved, this may be a blessing in disguise. A cynic might even argue that the evolution of weed resistance makes once highly successful herbicide-tolerant cultivars obsolete over time, thus clearing the way for new cultivars to enter the market and reducing the chance that an effective invention reaches the public domain as a generic cultivar after the end of the patent term. For a company like Monsanto, the emergence of glyphosate-resistant weeds at a time when its patents on glyphosate-tolerant crops are about to expire is definitely not something to be deplored (although company scientists had earlier dismissed this very possibility as highly improbable). This process of creative destruction favors private "innovation." Industry scientists claim that the use of new transgenic crops with stacked tolerance traits for glyphosate and other herbicides like 2,4-D and dicamba is not likely to accelerate the evolution of multiply resistant weeds, but other researchers argue that sooner or later the emergence and spread of such superweeds is precisely an outcome that is to be expected (Mortensen et al. 2012). The whole agricultural system seems to be set on "transgene-facilitated herbicide treadmill" (ibid., p. 83). Unfortunately, the knowledge structure needed to practice integrated weed management, which would enable farmers to escape from this treadmill, is simultaneously atrophying, because the relevant type of

knowledge does not lend itself to being packaged in patentable and salable products (ibid., pp. 81–82).

Another example of "nature fighting back" is provided by the use of insect-resistant *Bt* cotton in China, which for a series of seven successive years brought lower spraying costs and improved health to Chinese farmers, until in the 8th year a formidable resurgence of "secondary pests" necessitated a much greater and very costly use of previously abandoned pesticides, completely eroding the advantages of *Bt* cotton (Wang et al. 2006). A final example is provided by South Africa, where the African maize stem borer (*Busseola fusca*) eventually developed such widespread resistance to Monsanto's *Bt* maize expressing the so-called Cry1Ab gene (MON810), that the cultivation of this hitherto extensively grown food staple variety had to be abandoned in 2013. Maize farmers in South Africa now pin their hopes on "stacked" varieties that combine the Cry1A.105 and Cry2Ab toxin-producing transgenes, but it might be just a matter of time before pest resistance to these new varieties emerges, especially when appropriate pest management strategies (like maintaining refuges planted with non-*Bt* crops) are not complied with (Van den Berg et al. 2013).

## Design for Values?

The foregoing discussion shows that many different values may be at stake in the development of agricultural biotechnology. The question is whether value-sensitive design or design for values can be said also to play a part in this field of technology. Insofar as agricultural biotechnologists deliberately try to "build" certain "traits" into existing crop varieties, their work can undoubtedly be described as a form of design. Yet there are some difficulties that would militate against a straightforward application of value-sensitive design.

One major complication is that this "building" activity occurs mainly at the *molecular* level of DNA, while the intended traits represent *phenotypic* properties of entire organisms that are also dependent on environmental conditions and genetic background (other genes already present). At first sight it might appear hardly deniable that a gene construct derived from *Bacillus thuringiensis* that codes for the production of an insecticidal toxin actually confers the trait "insect resistance" to the plants in which it has been inserted, but this holds only so long as the target insects have not become immune to the relevant toxin (and also, of course, on condition that the *Bt* transgene will be "expressed" in the plant in sufficient quantity). Thus Monsanto's *Bt* maize based on the Cry1Ab gene no longer protects, as it once did, against the African maize stem borer after the insect developed resistance against the Cry1Ab toxin. With the complex traits involved in various forms of abiotic stress tolerance, the links between genes and traits are even more tenuous and complicated. This is actually a major reason for critics to cast doubt on the expected performance of drought-tolerant GM crops, especially as long as the desired trait is created through the insertion of a single gene. The underlying issue here is whether and to what extent "biology" is indeed amenable to "engineering."

The fact that, say, a new insect-resistant crop can only be temporarily successful but will normally not be a durable innovation, marks a characteristic feature of technological design in the biological domain. The new crop does not simply become obsolete due to further technological change; its untimely depreciation is as it were biologically preordained. In 1973, at a time of rising environmentalist awareness, some German philosophers of science already criticized the scientific-technological view of nature as an infinite reservoir for technical intervention and the concomitant assumption of the endless reproducibility of experimental effects. They used an interesting example to make their point: "The validity of the claim that the chemical substance DDT has an insecticidal effect is warranted by a reproducible experiment. Actually, the experiment has been repeated millions of times, albeit not in the laboratory but through the technical application of DDT. But precisely this large-scale repetition invalidates the claim that DDT is an insecticide, as the massive use of DDT leads to the selection of resistant insect strains" (Böhme et al. 1973, pp. 141–42). This peculiarity of technological design in the biological realm is obviously highly relevant with regard to the *sustainability* of our innovations. We can ill afford to prematurely exhaust the limited natural arsenal of *Bt* toxins by developing GM crops that set up selection pressures accelerating the appearance of resistant insect strains. The fact that much biotechnological innovation represents "perishable knowledge" also undermines the rationale of intellectual property protection. In the standard account of the fictitious contract that is concluded between an inventor and society (the so-called patent bargain), the inventor who discloses his invention by giving a full description of it receives in return the exclusive right to use his invention for a limited period of time. After the expiration of the patent, his invention is supposed to fall into the public domain, that is, it has to be made freely available to society. If, however, the invention "perishes" in the course of the protection period, society will in the end see itself robbed of its part of the bargain. The problem of the "vanishing public domain" is not only relevant for agricultural biotechnology; it also plays a prominent role in the development of new antibiotics and the preservation of the usefulness of existing antibiotics (Outterson 2005). It would seem that the problem calls for some institutional redesign of the system of intellectual property.

The fact that agricultural biotechnology has predominantly been developed in a commercial setting also helps explain the virtual absence of any serious design for values. In the past 20–30 years, most biotech applications have been designed with *agronomic* traits in view. To make GM seeds attractive to farmers, they must offer benefits like increased yields, more resistance to insect pests, or reduced labor needs. The introduction of herbicide-tolerant and insect-resistant GM varieties clearly made sense from this perspective. The development of herbicide-tolerant crops that could in particular withstand the company's own proprietary herbicide (Roundup in the case of Monsanto) was also economically smart: it allowed Monsanto to make a relatively smooth transition from an agrochemical to a biotech company. Roundup-Ready soybean, maize, and canola also brought environmental benefits, especially because they stimulated low-tillage or zero-tillage agriculture, although these benefits had not been originally designed, but resulted

from farmers' initiatives. One could argue that it would have been morally better to develop new varieties that would not need any herbicide spraying at all, but in the absence of evidence that such options were really within the available design space, the argument remains rather hypothetical. There is a lot of historical contingency in innovation.

It does not always make sense to represent the work of agricultural biotechnologists as if it occurred in some abstract "design space." Such a design space, if it existed, would ideally delineate the various possible combinations of "traits" in plants that can be realized with the available tools of the trade at hand and would also suggest ethically relevant "trade-offs" between different traits (insofar as these traits can be linked to important values). However, in cases where complex traits are under consideration, as with various forms of abiotic stress tolerance, the links with the multiple relevant genes are so complicated that the idea of a clearly circumscribed design space loses its analytical utility. Even a seemingly simple single-gene trait like insect resistance starts to look more complex once we also take into account the likely indirect effects of its prolonged large-scale use on the evolution of the target insects.

Although the alleged inadequacy of a single-gene approach for tackling complex traits is prominently cited by critics to cast doubt on the promises of drought-tolerant GM crops, this point of criticism is also partly accepted by some of the proponents. Thus an otherwise favorable report on drought tolerance in maize comments: "Drought tolerance is a genetically complex trait, so it is reasonable to expect that a successful transgenic strategy will rely on transcription factors and cascades of genes, or transformation with several transgenes affecting different but key processes. *However, current attempts appear to be focused on single genes*" (Edmeades 2013, pp. 20–21; my italics). Here it is admitted that the current biotech approach falls short of what is considered the ideal strategy. What is described as such ("cascades of genes … affecting different but key processes") actually looks quite similar to what is also known as *metabolic engineering*, whereby entire biochemical pathways controlled by networks of concatenated genes are being installed in a host organism. Metabolic engineering is an important set of tools for the emerging field of synthetic biology. A famous example is the creation of a complete new biochemical pathway, controlled by 12 genes from three different organisms, in yeast cells for the production of a precursor of artemisinin, a medicine against malaria, by Jay Keasling's team in Berkeley, California – a landmark achievement that figures as a poster child for synthetic biology.

Synthetic biology is described by many of its practitioners as the attempt to make biology (more) easy to engineer. In their eyes, what passes for "genetic engineering" in classical biotechnology hardly deserves this term at all or can only be considered a very primitive form of engineering. Critical NGOs like the ETC Group, by contrast, tend to portray synthetic biology as "extreme genetic engineering," thus emphasizing the continuity with biotechnology. It is useful to keep in mind that no clear dividing line can be drawn. Synthetic biologists aim to create standard biological systems from standard devices which in turn are

produced from well-characterized standard parts. Their designs have to satisfy the engineering requirement of modularity, so as to ensure predictable performance when different parts are assembled together to form a new system. Currently, the program of synthetic biology is more a promise than a reality, although many synthetic biologists all over the world are trying hard to turn the promise into a reality. The success of a new paradigm, as Thomas Kuhn already famously noted, is "at the start largely a promise of success" (Kuhn 1970, p. 23). Only time can tell whether synthetic biologists will ultimately succeed in effectively taming unwieldy biological complexity. It is clear, however, that this effort is confronted with huge challenges (Kwok 2010). Practitioners respond differently to those challenges. Some reaffirm their confidence that the new field will indeed rise to the occasion (Kitney and Freemont 2012), while others profess the value of humility in the face of the overwhelming complexity and unpredictability of the biological world (Agapakis 2014). The dominant attitude in synthetic biology is arguably still one of Promethean overconfidence, as testified by the frequently repeated claim that the range of useful applications that synthetic biology potentially holds in store is "only limited by our imagination." This attitude may boost confidence but is not conducive to a serious evaluation of the moral dilemmas to which the application of synthetic biology may give rise. It easily leads to a denial of all constraints, so that everything is possible and no hard choices need to be made between different ends. This is not to deny that technology, including biotechnology and synthetic biology, may relax existing constraints and thus help to create room for striking more acceptable trade-offs between competing values. What offers grounds for hope is that serious attention to ethical and social aspects of new applications forms an integral part of the international iGEM International Genetically Engineered Machine student competition, which works as a training ground for attracting new recruits to synthetic biology. Practitioners may thus gradually learn to overcome their overweening confidence. In Europe, finally, the wish to avoid another GMO debacle (the rejection of GMOs by a large part of the population) is a strong motive for policymakers to support initiatives for what is currently called Responsible (Research and) Innovation. This too affects the social matrix in which synthetic biology will evolve.

In the near future, gene technologists may become more fully aware that "design for values" is the name of the game. One major technological challenge for biotechnology and synthetic biology is to solve the harsh dilemma of "food *or* fuel," which gained visible prominence by the backlash of higher food prices and deforestation that followed the first wave of enthusiasm for "first-generation" biofuels. More advanced (second, third, or fourth) generations of biofuels, based on various foreseeable breakthroughs and milestones, are expected to loosen up or even overcome the trade-off between the two major competing uses of the world's biomass. Such expectations could be no more than merely the beginnings of another cycle of hope and hype (Bindraban et al. 2009), but the adherents of biotechnology and synthetic biology are convinced that the dilemma is ultimately going to be solved, so that in the end we can have our fuel and eat it too (Graham-Rowe 2011). It would make an excellent test case for design for values.

## Cross-References


- ▶ Conflicting Values in Design for Values
- ▶ Design for the Value of Sustainability


## References


African Centre for Biosafety (2013) Africa bullied to grow defective Bt Maize: the failure of Monsanto's MON810 maize in South Africa. African Centre for Biosafety, Melville

Agapakis CM (2014) Designing Synthetic Biology. ACS Synthetic Biology 3(3):121–128

Banning C (2012) Restafval van planten als alternatief voor aardolie. NRC Handelsblad, 2 Mar 2012

Bernauer T (2003) Genes, trade, and regulation. Princeton University Press, Princeton

Bindraban PS, Bulte EH, Gordijn SG (2009) Can large-scale biofuels production be sustainable by 2020? Agr Syst 101:197–199

Böhme G, van den Daele W, Krohn W (1973) Die Finalisierung der Wissenschaft. Zeitschrift für Soziologie 2(2):128–144

Bowman v. Monsanto Co et al (2013) No 11–796, slip op (S.Ct. 13 May 2013)

Carson R (1962) Silent spring. Houghton Mifflin, New York

Charles H, Godfray J, Beddington JH, Crute IR, Haddad L, Lawrence D, Muir JF, Pretty J, Robinson S, Thomas SM, Toulmin C (2010) Food security: the challenge of feeding 9 billion people. Science 327:812–819

Correa CM (2006) La disputa sobre soja transgénica: Monsanto vs. Argentina. Le Monde Diplomatique/El Dipló, Apr 2006

Darwin C (1972 [1859]) The Origin of Species. Penguin Books, Harmondsworth

De Schutter O (2011) The right of everyone to enjoy the benefits of scientific progress and the right to food: from conflict to complementarity. Hum Rights Q 33(2011):304–350

Doorman M (2012) Rousseau en ik. Bert Bakker, Amsterdam

Edmeades GO (2013) Progress in achieving and delivering drought tolerance in maize – an update. ISAAA, Ithaca

Glover D (2009) Undying promise: agricultural biotechnology's pro-poor narrative, ten years on. STEPS working paper 15. STEPS Centre, Brighton

Graham-Rowe D (2011) Beyond food versus fuel. Nature 474:S6–S8

GRAIN (2007) The end of farm-saved seed? Industry's wish list for the next revision of UPOV, GRAIN briefing, Feb 2007, Barcelona

Grant H (2008) Our commitment to produce more, conserve more. http://www.monsanto.com/newsviews/Pages/OurCommitmenttoProduceMore,ConserveMore.aspx

Griffiths P, Stotz K (2013) Genetics and philosophy: an introduction. Cambridge University Press, Cambridge, UK

Gurian-Sherman D (2012) High and dry. Why genetic engineering is not solving agriculture's drought problem in a thirsty world. Union of Concerned Scientists, Cambridge, MA

Harhoff D, Régibeau P, Rockett K (2001) Some simple economics of GM food. Econom Policy 16 (33):265–299

Herring RJ (2007) Stealth seeds: bioproperty, biosafety, biopolitics. J Dev Stud 43(1):130–157

IAASTD (2008) Synthesis report of the international assessment of agricultural science and technology for development. Washington, DC. http://www.agassessment.org/

James C (2012) Global status of commercialized biotech/GM crops: 2012, vol 44, ISAAA brief. ISAAA, Ithaca

Jasanoff S (2005) Designs on nature: science and democracy in Europe and the United States. Princeton University Press, Princeton

Keim B (2012) New GM crops could make superweeds even stronger. Wired, 1 May 2012

Kilman S (2010) Superweed outbreak triggers arms race. Wall Street J, 4 June 2010
Kitney R, Freemont P (2012) Synthetic biology – the state of play. FEBS Lett 586:2029–2036
Kuhn T (1970) The structure of scientific revolutions. The University of Chicago Press, Chicago
Kwok R (2010) Five hard truths for synthetic biology. Nature 463:288–290
Mittler R, Blumwald E (2010) Genetic engineering for modern agriculture: challenges and perspectives. Annu Rev Plant Biol 61:443–462
Mortensen DA, Egan JF, Maxwell BD, Ryan MR, Smith RG (2012) Navigating a critical juncture for sustainable weed management. BioScience 62(1):75–84
Outterson K (2005) The vanishing public domain: antibiotic resistance, pharmaceutical innovation and global public health. Univ Pittsbur Law Rev 67:67–123
Rousseau JJ (1966 [1762]) Emile ou de l'éducation. Garnier-Flammarion, Paris
Scoones I (2002) Can agricultural biotechnology be pro-poor? A sceptical look at the emerging "consensus"'. IDS Bull 33(4):114–119
Vaidyanathan G (2010) A Search for regulators and a road map to deliver GM crops to third world farmers. The New York Times, 31 Mar 2010
Van den Berg J, Hilbeck A, Bøhn T (2013) Pest resistance to Cry1Ab *Bt* maize: field resistance, contributing factors and lessons from South Africa. Crop Prot 54:154–160
Wang S, Just DR, Instrup-Andersen P (2006) Tarnishing silver bullets: Bt technology adoption, bounded rationality and the outbreak of secondary pest infestations in China. Paper presented at American agricultural economics association annual meeting, Long Beach, 22–26 July 2006

# Design for Values in Architecture

Lara Schrijver

## Contents



**Abstract**

The notion of design for values, or value-sensitive design, is founded on the idea that design principles are related to ethical, moral, social, and political values. In architecture, a general relation between values and design is present throughout the history of the discipline. However, the question then arises *which* values are related to design principles and *how*. This chapter examines architecture as a general application domain in which values have been of central concern throughout its history. It departs from the supposition that values are by necessity part of the project of architecture and unravel aspects of these values. These aspects include the distinction between implicit and explicit values, the

L. Schrijver (✉)
University of Antwerp, Antwerp, Belgium
e-mail: Lara.Schrijver@uantwerpen.be



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_22

unexpected effects of design intentions, the distinction between general values and their particular (historical) readings, and perhaps most importantly the life-span of buildings, which often outlasts the value systems they arose from.

**Keywords**
Values • Ethics of architecture • Design for values • Architecture and morality

## Introduction

The notion of design for values, or value-sensitive design, is founded on the idea that design principles are related to ethical, moral, social, and political values. Emerging from pressing concerns on human interaction with technology, particularly in the domain of computer science, value-sensitive design takes human values and interaction with technological systems into account from the initial phase of design. While design for values is a convention that is more widely used in relation to technology and in fields such as industrial design rather than in architecture, one might argue that within architecture, value-sensitive design has been incorporated from its very beginnings. From the first documented reflections on architecture, on its role in and relation to society, human values and human interaction with the designed environment has been a central concern. Or, as Churchill phrased it: "First, we shape our buildings and then our buildings shape us." While this often used comment appeals to an intuition that our built environment has an influence, the precise nature of this influence is still uncertain. What is the relation between architecture, design values, human interaction, aesthetic principles, and ethical concerns? How do design values or architectural principles give shape to social or moral values? How might buildings embody moral principles, and how precisely are they articulated?

These questions have many extensions, not all of which will be included in this chapter. Instead, the core issues of the entangled relationship between human values, the built environment, and the symbolic ascription of values to buildings will be positioned historically and drawn out to concerns of today. In other words, to what extent does architecture not only reflect our values but also shape our behavior? How has this been seen in the past, and are there significant transformations to be found? In this chapter, architecture is taken as a general case of how values are inscribed in artifacts, how artifacts or buildings are understood to have effects on human behavior, and as such, shape commonly held values.

As such, this chapter will focus on the underlying attribution of social agency to artifacts and position it historically in relation to the hopes we hold for architecture's influence, both for society and in the sense of cultural production. In other words: what does a building "do" in the sociopolitical domain, and how does it do this? This begs the question of the embodiment of values; it is here assumed that buildings can "possess" or communicate values. This is followed by a historical overview of the social and political values attributed to architecture as a profession

and to its built works in the Classical and the Contemporary ages. In this overview, particular attention will be given to a turning point in the nineteenth century, when an explicit social agenda was introduced in architecture in the wake of industrialization and an increasing urbanization.

In order to understand the particular challenges in the domain of architecture, this chapter offers a number of perspectives on how values are embodied within spaces and buildings, and how these values have evolved over time. This also raises the question of our understanding of values incorporated in the built environment, which becomes eminently visible in the written treatises on architecture, which are to some extent at liberty to ruminate on ideal situations. The chapter thus also discusses the distinction between the material reality of what is built and the ideal reality of design intentions, which is traced back to the gap between general value statements and their particular expressions in a specific building for a specific site, client, and context.

The chapter thus traces its path through questions of ethical and aesthetic values (both of which are immanent to architecture) and of their reception – whether the city occupant or the building user is attuned to the values that architects and urban planners inscribe within the spaces they design. Current and recent research on constraints and affordances, on agency and action schemes, and on the tacit knowledge and values embedded in material artifacts stands as the breeding ground for future research in this area.

The concluding section will discuss a number of concerns on value attribution to objects, such as the question of how they influence behavior, whether intentional value inscription can be understood unequivocally, and which consensus may be found. This section thus will situate a number of potential future research questions in the domain of design for values.

## Architecture: The Spatial Embodiment of Values

Historically, architecture is understood to embody values on two levels. On the one hand, there is the unconscious embodiment of the accepted values of a society. On the other, there is the intentional inscription of values that the architect or patron believes *should* be held.

To begin with unconscious values: these are values that over time congeal into spatial "habits," such as placing the hearth at the center of a home or separating subsidiary circulation routes within a building. In the case of the hearth, there is an identifiable historical core that derives from the traditionally central place of cooking and warmth in a preelectric era. Over time, this kernel of practical concern has accrued the meaning of the warmth of the home, long after the functional necessity of the hearth disappeared. As to circulation routes within a building, there are homes with separate circulation routes that once served to allow invisible access to all spaces for the servants of the household. As such, these routes still bear implications of a class society that remains present in the spatial organization of buildings.

The second approach, in which architecture is deemed to not only encourage preferred behavior but also to shape our underlying value systems, is particularly pronounced in the modern age. As such, it shares features with avant-garde art, which envisioned the possibility of changing collective values through new forms of artistic production. This primarily modern understanding has its roots in the nineteenth century and understands architecture to not only guide our behavior, but in so doing, to shape our values. This forms the heart of nineteenth- and twentieth-century progressive architecture in which urban planning and architecture were seen as a manner to not only improve the built environment but also to encourage preferred forms of behavior. In this approach our buildings "act" and are not mere backdrops that set a scene in which social groups and individual urban occupants can show completely independent behavior. This approach is not generally accepted. There are those who believe buildings do not "act" but simply reflect dominant aesthetic principles or the most functional spatial solutions. There are others who see our buildings as equal partners in the formation of society and contribute fundamentally to how we as a society act. Most people see the merit of both criticisms; the built environment may have some influence on our behavior and values, but this influence is ambiguous. Additionally, it is not precisely determined; human occupants at times will go straight against the grain of the intended or implicit values.

One of the most provocative positions on the effects of the built environment is to treat it as an "agent" itself, following along the lines of actor-network theory. This position is founded on an increasingly strongly articulated hope for the emancipatory influence of architecture in recent history. It is part of a more than century-long development, in which the effects of architecture were envisioned to be extremely far-reaching and as having the potential to reconfigure society through presenting new types of environments. Actor-network theory in some sense tones down this extensive influence by suggesting that our buildings are one of many factors (Till and Schneider 2011). Yet it also accords a level of independent "agency" to the building, suggesting an impact far beyond being a mere backdrop. In other words, buildings do more than just "sit there." They may influence our moods, our behaviors, and over time indeed even our ideas and values.

As such, our buildings not only embody certain values within their very design, they "enact" or "suggest" certain ways of living – norms we may or may not hold to (van den Hoven 2013). They suggest in their very design certain actions, and as such they may *intimate* certain behaviors. Yet these are not clean-cut behavioral schemes (Illies and Meijers 2014: 165). As such, architecture operates within a spectrum of values that are embedded within design, yet may or may not have the presumed effect (Gans 1968). This discussion has been perhaps the most prominent concern in the twentieth century, as it ranges between the extremes of architecture's inability to change people's lives and at the same time also the knowledge that destructive planning projects can be devastating not only to the environment but also to the sociocultural fabric of a city. While there is growing consensus on the presence of essential values that are communicated within our material objects, it remains difficult to ascertain not only which precise values are communicated, but also how stable these values are.

Moreover, the implicit values may be ambiguous themselves. If a space has very large dimensions, and its effect is to make the occupant feel small, what does this mean in terms of human sensibility? In many churches or institutional buildings, scale is used to create a sense of 'something bigger," whether that refers to "God," "the public domain," or "authority." Yet one might also suggest that feeling small may make one feel incapable of making a difference.

At the same time, there may be very small and practical interventions that change how we use a space; a recent experiment encouraged people to take stairs instead of the elevator, simply by introducing yellow lines on the floor aiming at the stairs, gently "nudging" people into more healthy behavior. Should we applaud this experiment and explore how we might more substantially introduce healthy behavior in our buildings? Or does this say little about the values users of buildings hold and only shows how unconsciously they follow spatial triggers?

To understand these implications, we should first trace a path back through modern thinking in architecture. Modern architecture in particular aspired to an explicit influence on our collective values and sees a potential to shape them through aesthetics. This very hope for "social design," or the influence on the collective through aesthetics, has remained with us since the middle of the nineteenth century.

## Articulating Values in Architecture: Writing and Building

Architecture treatises, throughout history, have been a manner for architects to articulate their principles of design and their relation to societal and aesthetic values. Beginning with the earliest known surviving manuscript, *The Ten Books of Architecture* by Vitruvius, treatises on architecture offer a self-reflection on the role of the architect as well as specifying which fields of education or work might be deemed central to the profession. Although each treatise is the product of a particular individual view on architecture, as a body of knowledge, the treatises offer an overview of societal and disciplinary concerns at particular moments in history. This includes not only the explicit concern for the appropriate composition, form, and aesthetic principles of a building being addressed but also some of the values implicit in these formal expressions. The treatises should be read with the awareness that they were often written by practicing architects who also aimed at legitimizing their own work. Yet within these limitations, they articulate the values at stake at certain historical moments, the aesthetic forms deemed appropriate to architecture in relation to these values, and the individual position of the author.

In essence, it is difficult to speak of architecture *without* turning to the values embodied within, or at least referred to, as touchstones for the designs provided. While Vitruvius and the Renaissance architect-painter Leon Battista Alberti form the basis of the discipline, it is in the nineteenth century that the sociopolitical and moral values are most brought to the foreground. In particular, Augustus Welby Pugin and John Ruskin make an explicit appeal to morality in their support of Gothic architecture (Pugin 1836; Ruskin 1849; Kruft 1996: 327-329, 331-33).

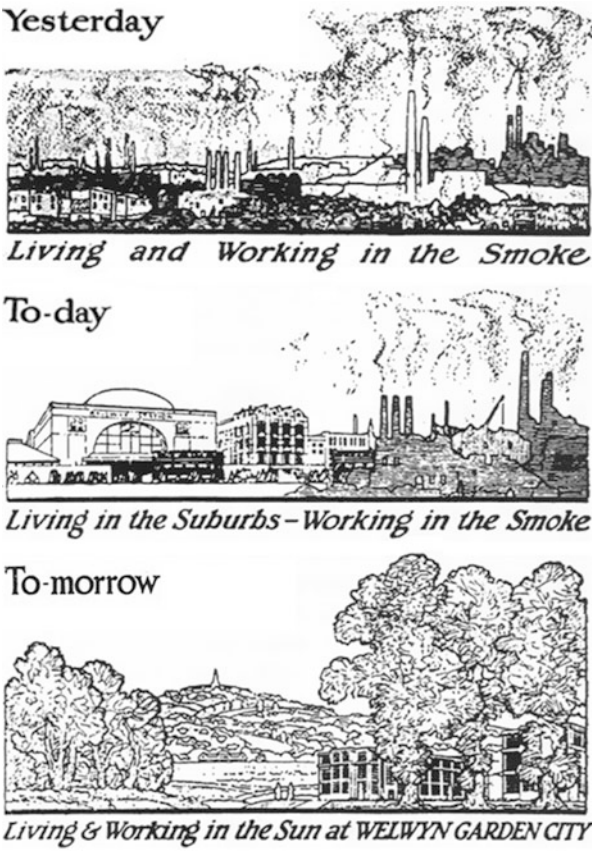



**Fig. 1** Garden City cover image

This is to be found throughout many different lines of thought in the nineteenth century. Similarly, ideals of social emancipation and improvement form the core of a number of (semi-)utopian schemes, many of them seeking improvements to the industrial city. For example, Ebenezer Howard's 1898 scheme for the Garden City limits the size of the city to 32,000, envisioning each garden city as a self-sufficient community. He organizes the city in a compact radical structure around parks and housing, preserving the countryside and limiting commuting time (see Figs. 1 and 2).

What these examples share is their support of particular (formal) principles of architecture with ethical arguments. In the case of Ruskin, the Gothic style of architecture was seen to embody the grace of natural growth, elegance of construction, as well as authenticity. Ruskin envisions these abstract moral values as implicitly present in the mode of Gothic construction and suggests that they encourage such values in the beholder. However, legibility is an unresolved issue here. Ruskin's treatises contribute to a specific reading of Gothic architecture, but the question remains whether this forms the only possible reading. The difficulty in all these positions remains the slippery foundations on which they are built.

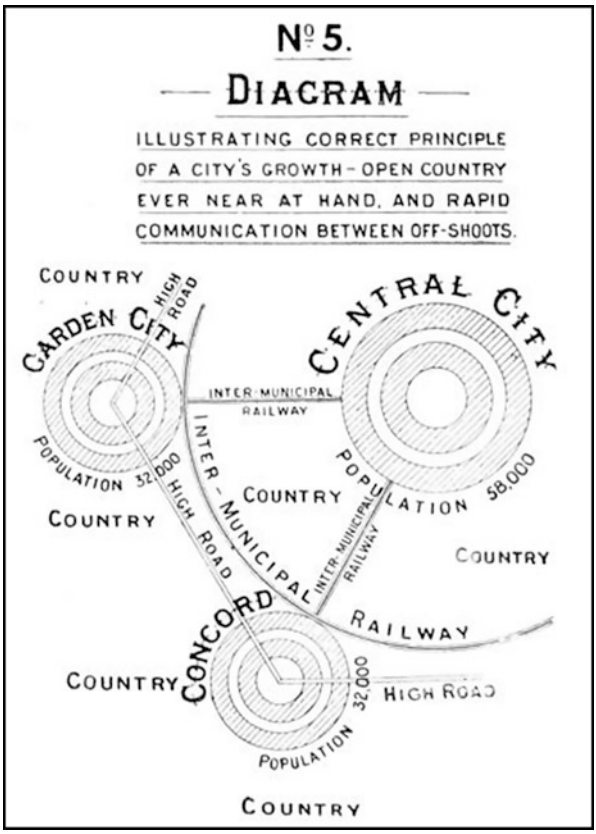



**Fig. 2** Garden City urban diagram

If one believes that Gothic cathedrals express an incontestably honest form of structural integrity, then one may perhaps follow the line of reasoning to emphasizing particular values. However, it requires an understanding of "honesty" that relates to the structural integrity of Gothic architecture, which then includes nonstructural additions and ornaments, fulfilling either a didactic or an aesthetic function.

As such, the values deemed an integral part of the architectural project are often also circumscribed by the specific debates in architecture and general values used to symbolize them. The very notion of "comfort," for example, might be interpreted in a markedly different manner, depending on whether one is describing a modernist home or a nineteenth-century interior. The modernist home might be founded more on a sense of airiness, light, and spaciousness – the freedom from clutter as a specific interpretation of "comfort" – while the nineteenth century might be more focused on the interior and on an enclosed intimacy (see Figs. 3 and 4). Moreover, the abstract notion of comfort, even if it could be taken out of a cultural context, might vary greatly in relation to other contextual conditions or depend on whether one addresses it as a technological issue or a design question.

**Fig. 3** A nineteenth-century interior

**Fig. 4** A modern interior

This is perhaps also why architectural treatises of times long gone continue to be used today. As we study historical styles, it is in writing that a number of the underlying suppositions seem to be more defined. One might speak of social cohesion, but is this exemplified by a building with a courtyard for all inhabitants or rather by an apartment building with a long gallery? As such, the treatises identify and situate commonly held (implicit) values and offer a translation into architectural form. The treatises, in essence, form a translation guide from general values to particular articulations.

## Values in Architecture: General and Particular

Throughout architectural treatises as well as project descriptions and critical evaluations of buildings, "big words" recur. Issues such as authenticity, spirit of the age, emancipation, and social progress are set at the center of architectural concerns.

These large, encompassing notions suggest that we indeed hold high hopes for what works of architecture may do. Or, as Paul Scheerbart suggests in his manifesto on *Glass Architecture*: "Our culture is in a sense a product of our architecture. If we wish to raise our culture to a higher level, we are forced for better or for worse to transform our architecture" (Scheerbart 1914: thesis I).

The modernists in particular make use of grand rhetoric, but it is already present in the earlier work of the nineteenth century. As issues of emancipation and social transformation become more central to architectural positions, the actual incorporation of the values identified becomes more visible in particular examples. For example, as modernist architecture aims at improving life for the masses, its aspirations hardly differ from nineteenth-century suggestions that architecture may ennoble or emancipate its inhabitants. At the same time, modern architecture is radically distinct from earlier designs as it makes use of a particular aesthetic founded on industrialism and rationality. For example, the 1926 Frankfurt Kitchen proposed by Margarete Schütte-Lihotzky makes use of industrial techniques and insights from Taylorism. The minimal dimensions limit unnecessary movements and make the space cost-efficient to produce. This in turn makes it available to the masses thanks to industrial production. The Frankfurt kitchen thus sets aside traditional values of "warmth" in favor of a modern sense of living, founded on efficiency of movement and a minimum dwelling standard for all.

In other words, modernist architecture reiterated general values of the "spirit of the age" or the "authentic" nature of man. Yet this was in a context of the late nineteenth century, which was marked by the industrial city. As such, concerns such as hygiene, fresh air, and light were substantiated in the clean lines of modernist architecture, a particular translation of "authenticity" as involving no unnecessary ornament. The aim of hygiene was also supported by its surfaces, which were less prone to invisibly collecting dust. The white walls of modernism as such produce corollary effects of requiring cleanliness, which in turn had unforeseen consequences in requiring a heightened attentiveness to the household (Wigley 2001; Lupton 1996).

By and large, it seems to have been the radical aesthetic innovations that most marked the legacy of modernism. Its proposition of a "new" architecture to accompany the spirit of the industrial age hoped to replace existing cultural values with new values envisioned for an era of industrialization. As such, it took a leap away from existing cultural perception and aimed at shaping new meanings. The modernist rhetoric of functionalism intentionally glosses over the cultural, aesthetic, and social connotations of design. While the actual, material transformations to the built environment were often still carefully designed and responsive to their context, modernism as a whole is strongly determined by this rhetoric. John Haldane reflects on this as the loss of a self-evident sense of meaning in architecture, which is replaced by a willful construction of new meanings that is intentionally disconnected from existing cultural values. He suggests that in order to recover the cultural significance of architecture, we might look to the "premodern understanding of architecture as a domain of embodied meanings and values" (Haldane 1999, p. 9). His attempt to mend the divide between aesthetic and ethical functions is built

on the proposition that "architecture offers a particularly powerful refutation of the idea that aesthetic value is one thing and practical function another" (Haldane 1999, p. 9). This suggests a general interlacing of aesthetic value and use, intended or actual.

Haldane's appeal to a premodern understanding directs us to what seems a more self-evident relation between aesthetics and values in Classical and Renaissance architecture. And indeed here one might find that the general values specified in architecture treatises are relatively stable, but they differ in the specific forms they take. Values such as order, symmetry, and eurhythmy all figure in the work of Vitruvius already, yet they take on different forms depending on the time and context. "Order" can be interpreted as a fundamental value of both Classical and Gothic architecture, which however manifest their expression of order in quite different manners. Preferred proportional systems have varied to some degree over time and context, but the principle of proportion has been crucial in most architecture discussions, from Vitruvius to the treatises of the Renaissance and later, to the Modulor by Le Corbusier. "Ornament" is perhaps one of the most explicitly controversial categories since the advent of modernism (Loos 1908). However, even Alberti distinguishes between "mere decoration" and "ornament," which fulfills an aesthetic or perceptual necessity (Alberti 1452). Ornament is thus a crucial element in architecture, contributing to its aesthetic value. An implication of moral judgment is present in the distinction between frivolous or superficial "mere decoration" and aesthetic necessity. The central role of ornament has been raised again in recent work that explores more extensive possibilities for elaborate and customized ornamentation since the advent of digital fabrication (Spuybroek 2011; Picon 2013).

Similarly, while sociopolitical values in architecture have recurring general themes, they have undergone fundamental transformations in the material forms they take on over time. The very notion of "community," for example, returns time and again, most prominently in relation to urban compositions and institutional building. The inherent aspiration toward building a sense of community may be embodied in the Greek *agora* as a space for public discussion. Yet to a modern sensibility, the exclusion of women and slaves from public life seems too restrictive to merit the label "community." Many urban plans of the early twentieth century such as New Lanark Mills and the Garden City include specifically communal spaces. At the same time, the correlation of these values to a particular aesthetic is often weak. Is "community" best articulated by large public squares or by intimate urban streets? By accessible institutional buildings or by a network of smaller public spaces? Similarly, a shared value of "justice" might lead to Vitruvius' assertion that a dwelling should reflect the social status of its owner (Vitruvius, 27 BC), while in the postwar welfare state, a similar value might be deemed more adequately articulated in the egalitarian housing blocks of northwestern Europe (Mattsson and Wallenstein 2010, pp. 17–19).

In other words, general value assumptions seem relatively stable throughout fundamental changes of design principles, while the particular are more clear but easily susceptible to changing habits. This may be attributed to the openness of

general value assumptions. In contrast, particularly circumscribed values may contain the general appeals but are dependent upon specific understandings of these general notions.

At the same time, if architecture is indeed a domain of embodied meanings and values, it is in the nineteenth century that a great transformation takes place. In Classical architecture, the particular expressions of value are situated within an accepted aesthetic frame (of Classical architecture), and their meanings are in-line with social convention. In the nineteenth century, as the arguments over aesthetic principles are shifted to the domain of morality, a break appears between aesthetic form on the one hand and implicit ethical principles on the other (Watkin 1972).

## Morality and Aesthetics: Is the Good Always Beautiful?

The moral values actively presented within architecture of the nineteenth and twentieth centuries fulfill three roles within the debates. First, they form a foundation for aesthetic principles. As such, they provide a nonaesthetic argument for principles that might otherwise be deemed as arbitrary or subjective opinions. This is in part correlated to a diversification of aesthetic principles in the nineteenth century. As the language of classical architecture was no longer seen as the only legitimate design principle, fierce, stylistic, and aesthetic debates arose, in which most sought to prove their uncontested legitimacy. Second, moral assertions also provide support for the value of architecture in general. In other words, the practice of architecture is legitimized by its social impact. Finally, they also justify guidelines and building regulations that delineate minimal acceptable standards from the perspective of collectively held values. As such, they circumscribe the maximum constraints or minimum affordances to be incorporated within the architecture, such as accessibility or daylight.

While moral principles are noted in premodern treatises of architecture, they are typically limited to the conduct of the architect. In other words, while the architect could be held responsible for his professional demeanor and his integrity, this in itself was not reflected in the stones of his buildings. At the same time, as Vitruvius argues, this moral conduct is crucial, "for no work can be rightly done without honesty and incorruptibility" (Vitruvius: Bk I, Ch I).

In contrast, it is in the nineteenth century that these moral assertions become attached to architecture itself. Pugin first paved the way for the nonaesthetic valuation of architecture in *Contrasts* (Kruft 1996, pp. 327–329). To him, Gothic architecture in itself contained spiritual qualities that could be experienced in the light colored by stained glass and in the breathtaking height of the cathedrals. As such, Pugin shifted the perceived value from an aesthetic one to one of propriety to social values, paving the way for an increasingly socially oriented understanding of architecture. Or, as Fil Hearn suggests, Pugin "awakened the notion that good architecture, Gothic or otherwise, could both embody and reinforce social virtue" (Hearn 2003, p. 12).

These developments in the nineteenth century lay the groundwork for what we today see as the main "movement" to enforce particular social values within its designs: early modernism. Modernist architecture positioned social purpose as a central concern, aiming to incorporate particular values such as openness and transparency within the very design and structure of the architecture they built. Throughout the first half of the twentieth century, the potential for architecture (as indeed, for many of the arts) to transform life was placed at the center of the discipline. The work of Le Corbusier as well as that of Ludwig Mies van der Rohe and Walter Gropius can hardly be understood without taking into account the impending transformation of society they envisioned through the impact of industrialization and new technologies. Architecture was to be an aid not only in revealing the qualities of this new, clean, rational life but also in shaping it.

In the 1960s, Herbert Gans referred to this as the "fallacy of physical determinism" (Gans 1968). With this phrase, he took to task the many urban and architectural propositions that assumed a behavioral response correlating to the intention of the architect. The phrase itself identifies the remarkable conflation of moral and aesthetic values in the nineteenth and twentieth centuries. This is a widespread response to societal changes. As social conditions transformed rapidly under the influence of industrialization, the stability of sociopolitical conventions dissipated. The (self-proclaimed) role of the arts in general and architecture in particular became one of providing new meanings in a period of transformation. Groups as diverse as Dada, surrealism, and Russian constructivism set forth artistic principles that were deemed in accordance with new societal values originating from the Machine age. With the advent of modernism, seminal works such as Georg Simmel's "Metropolis and Mental Life" and Adolf Loos' "Ornament and Crime" began to define a body of work that interlaced aesthetic and ethical concerns at the outset.

This problem goes to the heart of the issue of values in architecture. Architectural practice makes use of a wide range of value assumptions. This is related to its inherent connection with human concerns, such as spatial use, privacy, domestic comfort, and similar issues. Architecture is after all integrated in a larger sociopolitical context yet also forms the background to the everyday functions of life (Lagueux 2004). It is also embedded within the history of the discipline, in which aesthetic principles and moral concerns are entwined and sometimes used to legitimate design premises. At the same time, the conscious application of aesthetic principles is seen as integral to the character of architecture. As such, the two domains are by definition unalienable from architecture. At the same time, the line between addressing necessary ethical issues and legitimating arbitrary aesthetic preferences is thin. Watkin argues that the treatises on architecture that have introduced particular arguments on morality, be it on the foundations of religion, politics, the Zeitgeist, or technology, have misrepresented architectural arguments by appealing to moral concerns (Watkin 1977, pp. 1–3, 13). As such, they base an individual aesthetic preference on attributions of justness or propriety.

It is clearly not always easy to distinguish between the two categories of evaluation. While the aesthetic appeals mainly to the domain of artistic creation, it is by the necessity-incorporated human occupation that simultaneously delineates

a domain of sociopolitical considerations and ethical concerns such as responsibility, security, or community. At the same time, considerations that arise out of architectural traditions and a concern for aesthetic expression have at times also triggered the reconfiguration of accepted values.

In the latter part of the twentieth century, values held high by modernism such as speed, dynamism, and ephemerality were countered by the "grounding" of human life in the built environment. A renewed concern for contextualism arose, as well as the importance of place in the *genius loci* and ideas of phenomenology (Norberg-Schulz 1979; Pallasmaa 1996). Most of these positions in some manner addressed the alienation triggered by modern architecture. Karsten Harries, for example, argues that the ethical function of architecture is to provide an interpretation of our time in the broad sense of its *ethos* (Harries 1997, pp. 2–13). In essence, he suggests that a permanent ontological discomfort that arises from the experience of modernity might be alleviated, if the art of architecture were again to engage with the task of providing a home or a place in the world. Harries draws on the work of Heidegger to argue that we should see the function of architecture as extending beyond the modernist, functionalist, or formalist perspectives to encompass a relation to the community and its sociopolitical fabric. This broader understanding continues to underscore the impact and the responsibility of our built environment while again leaving some of the specific articulations of this ethos open to interpretation.

## Value Attribution: Conduct or Object?

Building on the proposition that ethical and aesthetic judgments are naturally intertwined in architecture, the attribution of moral values then may also be extended from the conduct of the architect – the main domain of moral values in the treatises of the Renaissance and of Vitruvius – to also being implicitly present in design intentions, and residing within the objects of architecture. This is perhaps still implicit in Alberti's notion of *concinnitas*, "appropriateness," where the aesthetic value of a space is determined by the contextual factors of what it is meant to convey, and is thus delimited by the values of the society it serves. The explicit references to values in Vitruvius and Alberti focus particularly on conduct and such concerns as diligence and incorruptibility. The notion of "doing good work" thus transfers naturally to the work being "good." In other words, the principle of diligence is founded on an assumption of natural correlation. This assumption functions within a strong aesthetic framework of architecture principles, based on a clearly defined system of composition, proportion, and order.

The severing of a naturalized correlation in the nineteenth century leads to the issues raised by Haldane, when he suggests we redirect our attention to a more self-evident construction of meaning and value. Here, one might take issue with his characterization of modern architecture. In point of fact, the principles of modernist architecture clearly contain a consciousness of these embodied meanings and values. The attempt to transform the behavior and also the value systems of its occupants through architecture is founded on the supposition that the implicit value systems will be incorporated through their spatial presence. This is a noteworthy

development in that it inverts an earlier form of transference to its own advantage. As "doing good work" became equated with the work "being good," thus the inversion was also held to be true. Making architecture that "is good" will encourage those within to "do good" according to the principles implicit in the building.

This argument would have hardly been possible without the nineteenth century treatises that shifted value attribution from the conduct of the architect to the object of architecture. Pugin's rallying cry to social progress introduced a new spectrum of value judgments, which sees a form of didactic or a formational role for buildings. This is particularly taken up into the architecture debate with John Ruskin's *Seven Lamps of Architecture*, in which, for example, he argues by analogy to human behavior that false materials are morally reprehensible. The argument follows along the lines of the "white lie" told to a friend, which then casts doubt upon all previous and future statements (Ruskin 1849, Chap. II, Sect. I). Similarly, the effect of wood painted to look like marble or in the case of the British Museum, he addresses, a false granite "is to cast suspicion upon the true stones below, and upon every bit of granite afterwards encountered" (Ruskin 1849, Chap. II, Sect. XVI).

Ruskin categorizes architectural "deceits" in three types: structural, falseness of representation, and machine-made ornament. In the first two categories, the building essentially is seen to tell a lie. The building may appear to be held up by a series of structural columns, which are merely decorative rather than functional. Or, as in the granite of the British Museum, the material is masquerading as something else. In the third category, Ruskin implicitly returns to the mutual relation between object and conduct. In the case of machine-made ornament, the deceit lies less in representation. Rather, it is about the care and individual attention that handmade ornament contains, which a machine-made ornament cannot. This reintegrates object and conduct by seeing them as inseparable.

The object or conduct distinction situates judgment in an opposition between content and form, that is, something with horrific social content cannot be judged on aesthetic merit and vice versa. In architecture this complete severing of the two poses problems, yet a direct correlation equally remains problematic. Returning to the inherent presence of both aesthetic and ethical values within the domain of architecture, more sophisticated interpretations are currently emerging.

## Reweaving Values and Forms: Constraints and Affordances

Architecture may be taken as a domain of constraints and affordances that affect behavior but do not determine it. Affordances in architecture may suggest a certain action, a manner of using the space. At the same time, these affordances do not delimit human action to *only* this intention. In the simplest of terms, a brick wall with a window will likely lead the occupant to enjoy the view from the existing window. Nevertheless, it will not prevent the occupant from deciding to break a hole in the wall to seek out a different view. The existing configuration is relatively guiding, yet not fully deterministic. In this sense, we may see architecture as providing "action schemes," which tend toward preferred actions but do not inhibit

**Fig. 5** Still from *Mon Oncle*, 1958, Jacques Tati, kitchen scene; https://www.youtube.com/watch?v=LE9t98Gox60

other, less preferred actions (Illies and Meijer 2013). By focusing on these affordances as suggestive but not deterministic, it becomes possible to speak of values intentionally inscribed in the space without implying that these values are legible or indeed the only possible understanding.

To return to the Frankfurt kitchen, for example, the modernist functionalism that informs its design is focused on an efficient and rational use of space. As such, the kitchen is not a place to gather in but rather as a place in which food is prepared in the most efficient possible manner, with the least possible movements. The reduction in space and potential mass production cuts costs, thus making a completely furnished kitchen available to a larger percentage of the population. This offers a higher standard of living for more people, as intended. The limited space available, however, requires a very precise spatial arrangement, making it less readily adaptable to changing conventions or even (ironically) to developing technologies.

In the 1958 film *Mon Oncle*, Jacques Tati shows what happens when one does not follow the inscribed affordances of the functional modern kitchen. One scene shows him trying to get a water glass out of the cabinet in his sister's highly modern kitchen. He struggles with the kitchen's technology, and confronted with the self-opening kitchen cabinets, he drops the pitcher he finds in a cupboard. Fortunately, it simply bounces, offering a new unexpected twist in this representation of the functional environment. In this eminently humorous portrayal of modern life, Tati shows how crucial the understanding of implicit values is and what their unintended effects may be (see Fig. 5).

## Design for Values in Architecture

### Existing Approaches/Tools

As the main body of work that attributes moral values to design or attempts to incorporate values into the design arises from the nineteenth and twentieth centuries, it is also primarily in these examples that the explicit insertion of social and

moral values is to be found. The political drive in the late nineteenth century focused on making more livable cities and on improving dwelling conditions. This was not a luxury, as industrialization had driven a great migration to the cities, resulting in overcrowded dwellings, cramped living conditions, and a lack of proper ventilation. These conditions among others led to the work of Ebenezer Howard in proposing the Garden City as well as Frank Lloyd Wright's proposal for Broadacre City (Fishman 1982). As mentioned above, the Garden City proposed to limit the size of cities in order to mitigate the combined effects of industrialization and urbanization. The smaller urban centers of the Garden City would combine the amenities of the city center with the pleasures of outdoor living. Likewise, Broadacre City was designed to prevent overcrowding. It consisted of a large grid in which each family unit would occupy a half acre, which was deemed sufficient for sustenance. While the urgency of such societal concerns is immediately evident in light of increasing urbanization (how might we introduce acceptable dwelling spaces in rapidly overcrowding cities?), many of the additional features in these cities included a broad spectrum of values implicated in the design.

Many existing approaches that incorporate values in architectural design respond to similar urgent social concerns. This may include aesthetic concerns such as order, arrangement, proportion, and an overall "pleasing to the eye" appearance of the building. However, since the nineteenth century, the primary focus is on spatial composition as an explicit expression of desirable sociopolitical values. Influencing behavior through space thus underscores the moral values aimed at. Marking out particular gathering spaces, connections, and relations of the individual to the whole is a way of quietly inserting the basic principles of a social community into the very walls of its town.

Moreover, some of the ethical approaches take into account particular situations and historical contexts. If indeed privacy was a central concern to the second half of the twentieth century, the 2001 attack on the towers of the New York World Trade Center introduced a new consciousness of risk and security that pushed certain concerns for privacy to the background. However, as the extent of NSA surveillance is becoming increasingly apparent and the terrorist threat is receding into the background, privacy is again beginning to take center stage.

## Comparison/Critical Evaluation

Throughout various design approaches, the incommensurability (or irreconcilability) of intention and reception becomes manifest. They draw attention to the free agents of their occupants, who are indeed influenced by the spaces around them, and draw certain value appraisals from them but are also remarkably resilient in the insertion of their own value systems where the built environment is seen as inadequate or incommensurable. Therefore a certain consciousness of the temporality of these evaluations is helpful – the humility of knowing that what we now know to be true may change as our societies, environments, and insights change. This is the greatest difficulty in arguments on design for values.

While it is based on a strong analytical framework, its perspective is often focused primarily on the perspective of design intention or presumed use. The longer horizon of architecture and historical transformations in use, reception or understanding of our buildings, and the values that construct them also shows the unexpected, the innovative, the failure of certain "inevitable' successes, and the unexpected success of objects that were seen as doomed.

## Examples: Values and Transformation Over Time

Architecture has a long history, much of which may be traced back in its documents, its treatises, and its buildings. While we cannot always know for certain which values were predominant in the design process, we can historically evaluate public or critical reception, and we can identify moments of transformation in use or perception. As such, architecture contains a wealth of information for understanding values and design.

In the eighteenth century, French enlightenment architects such as Etienne-Louis Boullée and Claude-Nicolas Ledoux believed that pure and symmetrical geometric forms were the most adequate spatial translation for the emerging values of equality and universal rights. In the early nineteenth century, Jean-Nicolas-Louis Durand and Quatremère de Quincy were still hailing principles that could be applied in many different places. They continued to aim at a universal logic of building types and models. In the nineteenth century, Eugène Emmanuel Viollet-le-Duc and John Ruskin viewed the classicist design principles of such architects as oppressive, representing structures of political authority with heavy handed proportions. They sought instead to find a more genuine and individual relation to architecture, aiming at an architecture that was informed by its context (Kruft 1996, pp. 274–287).

The history of architecture offers a number of these examples, in which values once expressed with great conviction in specific built form, later stand symbol for something quite different. In 1933, a group of architects convening regularly with Le Corbusier as one of the key figures, proposed a plan for the functional city. The *Congrès Internationaux d'Architecture Moderne* (CIAM) thus proposed to separate the main features of the chaotic nineteenth-century city into distinct zones; separate dwelling, work, and recreation areas were linked by efficient traffic circulation. The accompanying text, the CIAM charter of Athens, suggested that not only would this rationalized city design be more attuned to a modern way of life but would encourage the city inhabitant to become more rational and efficient in behavior. As many cities were transformed in accordance with the CIAM plans, critics however called attention to the lack of distinction between one city and the next. The grid and the zoning principles, first hailed as perfect for the modern city inhabitant, were now seen as detrimental to the human use of urban space. One of the most broadly known critics of modernist urban planning, journalist Jane Jacobs, went directly against the grain of the functional city by reintroducing the mixed-use neighborhood. In her view, apartments above shops could ensure that there were

"eyes on the street"; someone would always be watching activities on the sidewalks and contributing to a sense of security. In addition, this reintroduced the potential of urban vitality through mixed-use, mixed-occupancy neighborhoods (Jacobs 1961, pp. 152–177).

Likewise, the modernist use of the glazed curtain wall was founded on the perception of glass as a material suitable to a new age of cleanliness, technology, and transparency. In the postwar institutional architecture of Europe, large-glazed buildings became common for offices and public institutions, underscoring transparency, both metaphorical and literal. Now, 60 years later, the global concern for dwindling resources "rereads" these buildings as wasteful of the energy they require to heat and cool.

As such, the duration of a building's life-span forms a different frame for understanding values inscribed in the design. As buildings are not always torn down for a new occupant, a space that has been intended for one type of occupation is often reappropriated quite differently. This requires a multilayered approach to values. There are those intentionally inscribed in design and others that may be added on over the course of a building's existence.

## Open Issues/Future Work

This component of time holds great potential for future work. Architecture, in contrast to many objects of utility, often outlasts the sociopolitical and cultural contexts it is realized in. Does the building of the Bauhaus in Dessau communicate the same radical innovation today than it did upon completion in 1923? One would surmise not, if for no other reason than the widespread distribution of some elements of the Bauhaus principles, whether in the sphere of artistic creation or in the catalogs of Ikea that are now omnipresent. Yet what remains of the Bauhaus is the consistency and strength of its architectural visions, the spacious halls, and the meticulous detailing. Is it comprehensible as a political statement in the context of today? Clearly not in the same depth, but perhaps it hints at its own context and ambitions.

The relation between architecture and social change is broadly supported in the current debate on architecture (see, e.g., the identification of architecture as both index of social change and as technology to deal with that change; Moore and Wilson 2013). While this is broadly held, the interpretations differ vastly, whether tending toward the functional, social, or aesthetic perspective on architecture. Moreover, the contemporary debate shows a rising interest in the role of these material objects and buildings as independent agents (Sennett 2008; Latour 2008; Van Eck 2009; Bierens 2013). Latour suggests a relatively radical understanding of a building as agent, taking the perspective of what the building "does"; "the way it resists attempts at transformation, allows certain visitors' actions and impedes others, bugs observers, challenges city authorities and mobilizes different communities of actors" (Latour 2008, p. 86). In his work on craftsmanship, Sennett discusses the importance of "resistance" both in shaping an outcome and in forming

the skill of the craftsman. Sennett notes, "Just as a carpenter discovers unexpected knots in a piece of wood, a builder will find unforeseen mud beneath a housing site" (Sennett 2008, p. 214). These contextual conditions of resistance "push back" at the craftsman, requiring an adaptation of the initial idea to material reality. Both Sennett and Latour lay a foundation for understanding the process of design and the resulting works of architecture and urbanism as a highly complex set of ideas, decisions, complicities, accidents, and responses to conditions at hand. This situates current work precisely between moral and aesthetic autonomy rather than at either end of the spectrum. Theories on system thinking and ecologies as well as the current turns toward materialism in the humanities take into account a more fundamental agency of material. The question is to what extent we can unravel particular features of these complex material-and-conceptual conditions in order to inform our design fields better.

As such, there is work to be done in the bridging of design thinking and architecture theories. The particular values utilized in everyday practice are to some extent codified in regulations, to some extent unconscious manifestations of cultural presuppositions, and at times are questioned through exploratory designs. It is in this domain that future work could set itself the task to be more precise in identifying general, weakly defined but strongly sensed values and more closely defined values that are easily susceptible to change. Finally, the key lies in the leap between understanding and doing, and how the values, which may be articulated in a design project, drive materializations that then transform the implicit assumptions we have.

In the meantime, seemingly trivial concerns such as floor height or particular proportions may prove to be of unexpected importance in the future. The Modulor, based on the relatively small dimensions of Le Corbusier himself, leads to relatively cramped spaces, in direct contradiction to the rhetoric of light and air he uses (Le Corbusier; Boyer 2012). In a time when northern Europeans are rapidly growing taller, the difficulty of the low door becomes a functional concern beyond the triviality of everyday.

This approach may add a specifically aesthetic perspective to the current work on design for values. As Van der Hoven notes, the work of value-sensitive design approaches the ethical dimension in order to understand more of its implicit assumptions:

> In value-sensitive design, the focus is on incorporating moral values into the design of technical artifacts and systems by looking at design from an ethical perspective. It is concerned with the way our acting in accordance with moral values (e.g., freedom, equality, trust, autonomy, privacy, and justice) is facilitated or constrained by technology [. . .]. Value-sensitive design focuses primarily and specifically on values and requirements of moral import. (van den Hoven 2013, p. 137).

Perhaps the rise of interest in the independent "life of things" that is so strongly evident in the art and architecture debates may continue on this trajectory of understanding moral assumptions while at the same time taking into account the multiplicity of interpretations embodied within the artifact.

Van der Hoven notes that ideas about values and morals get designed in. In many cases, these values eventually become codified in regulations, such as those on security and privacy, or minimal adequate living standards for social housing. Van der Hoven distinguishes these frameworks from those of functional import such as efficiency and storage capacity, yet the question might be raised whether moral and functional issues are easily to be distinguished in the built environment. Certainly the accepted insights of the postwar era, particularly those of poststructuralism and related approaches, have revealed a plethora of *implicit* moral values that are presented under arguments of functional efficacy or historical topicality. Many of these implicit moral values slowly adapt, while the material results of their incorporation in the built environment remain. Yet the reception of these values and our moral assumptions also transform to which the boulevards of Paris stand testimony. While they were originally meant to constraint possible revolutionary uprisings, they are not typically seen as a military intervention, enabling the shooting of cannons and the corralling of people. Their spatial expression – large, wide, spacious boulevards – were meant to underscore the grandeur of these developments, which may still be experienced today, though somewhat attenuated by the intensity of the traffic along them. The rationalization process that led to the scale of the boulevards is now less immediately tangible, being replaced by a more romantic notion of the *flaneur* wandering along the Paris boulevards.

It is unquestionable that there has been what Van der Hoven identifies as a "design turn" in ethics, making it relevant or significant to ask moral questions about a design. As such, general notions of well-being have been transformed into clearly defined design qualities. At the same time, there are often considerations that come to the fore only after the fact. Some of these issues may initially appear purely functional but as understanding increases become more founded upon moral considerations – take the example of street lighting. For crime prevention, more lighting is better. Electricity usage however becomes an increasing issue in an age of limited resources – so is extensive street lighting in a space not often used then a sign of the wasting of resources and the immorality of wastefulness? The choices then become to either generate electricity in a more sustainable fashion (wind energy, solar energy) or to engage in a new design proposal in which the limitations of resources are incorporated in the design itself.

Before the nineteenth and twentieth centuries, a specific understanding of community in Europe was often "designed in," which had to do with the respectability of institutions. To have a grand building was a symbol for the community as a whole. Yet at the time of the French revolution, this was more likely to be seen as a power-hungry symbol of an institution not supported by the people. As such, future research may find itself facing the challenge of envisioning various potential readings – the multiplicity of interpretations that one might envision and that have an impact on how these projects are seen.

Generating urban scenarios and architectural design projects may aid in sketching a coherent vision of a potential future in which certain values are of

central importance. Design for values is a growing field of research, and its insights are rapidly becoming indispensable to the many fields of design, from IT to architecture and urban design. Within this domain, it becomes increasingly important to understand the overall implications of values incorporated in designs. Each future vision springboards from the present, while its future visions, particularly in terms of coherence, inform us about unforeseen possibilities.

## Conclusions

Overall, the concern for both the implicit and explicit values in architecture has become more prominent. This may be attributed to the decreased confidence that the values we hold are universal, transferable, or even intersubjective. In the face of the postmodern conviction that individual values may not be subsumed under an overall argument that holds for all, and yet the sense that community is to be valued, treasured, and perhaps even reinforced, the debate on values takes on a new urgency. The importance of community seems accepted again after a period of emphasis on individualism, yet the form it should take is now a concern. All in all, though, each of these arguments and their architectural counterparts continue to demonstrate how fundamentally intertwined our sense of space and architectural design is with the values we understand them to imply.

As such, it is perhaps unsurprising that architecture takes into account – insofar as possible, given the limitations of the unpredictability of the future – "the political and morally relevant effects that designs, built structures, and artifacts may have" (Van der Hoven 2013). "Consciously or unconsciously, deliberately or inadvertently, societies choose structures for technologies that influence how people are going to work, communicate, travel, consume, and so forth over a very long time." (Winner 1986, pp. 28–29; as quoted in Van der Hoven 2013) – the same can be said for our buildings. One might argue that our buildings are a little less "defined" in the direction of a particular use or goal, but the long-term influence is undeniably present. The office building that is based on an open plan structure may be renovated with interior walls, but the primary structure will remain until the building is demolished to provide space for a new building. This makes post-occupancy evaluation a potential direction for future work, not merely in terms of utility, but also in terms of the full breadth of architectural concerns.

The future is open and interesting. Issues raised by Latour in his critique of science as well as those raised by Sennett in his work on craftsmanship both direct us to the absolute necessity for incorporating values beyond the quantifiable in our reflections on architecture, yet both take into account the unexpected reinterpretations that may take place when faced with the actual, material object of our desires. As such, architecture has a role to play in understanding design for values, and at the same time it has wise lessons to offer for when we take our presumptions of moral values too far.

## Cross-References

► Mediation in Design for Values
► Design Methods in Design for Values

## References

Awan N, Schneider T, Till J (2011) Spatial agency: other ways of doing architecture. Routledge, London
Bierens C (2013) De Handgezaagde Ziel. Essay 8. Mondriaan Fonds, Amsterdam
Bolle E (2000) Romanticism and rivalry. Archis 4:66–69
Boyer MC (2011) Le Corbusier: Homme de Lettres. Princeton Architectural Press, New York
Cohen J-L (2012) The future of architecture. Since 1889. Phaidon, London
Conrads U (1970) Programs and manifestoes on 20th-century architecture (trans: Bullock M). MIT Press, Cambridge, MA [orig. Programme und Manifeste zur Architektur des 20. Jahrhunderts (1964)]
de Botton A (2006) The architecture of happiness. Pantheon, New York
Filarete (1965) Treatise on architecture (trans: Spencer J). Yale University Press, New Haven [orig. Trattato di Architettura (ca. 1465)]
Fishman R (1982) Urban utopias in the twentieth century: Ebenezer Howard, Frank Lloyd Wright and Le Corbusier. MIT Press, Cambridge, MA
Förster K (2012) The housing prototype of the institute for architecture and urban studies. Negotiating housing and the social responsibility of architects within cultural production. Candide 6
Fox W (2000) Ethics and the built environment. Routledge, London
Gans H (1968) Urban vitality and the fallacy of physical determinism. In: Gans H (ed) People and plans. Basic Books, New York, pp 25–33
Haldane J (1999) Form, meaning and value. J Archit 4(1):9–20
Harries K (1997) The ethical function of architecture. MIT Press, Cambridge, MA
Hearn F (2003) Ideas that shaped buildings. MIT Press, Cambridge, MA
Hinte E (2014) Voorstel voor een nieuw accent. Archined. (http://www.archined.nl/opinie/2014/voorstel-voor-een-nieuw-accent/). Accessed 9 Apr
Illies C, Meijers A (2009) Artefacts without agency. Monist 92(3):420–440
Illies C, Meijers A (2014) Artefacts, agency, and action schemes. The moral status of technical artefacts. Philos Eng Technol 17:159–184
Jacobs J (1961) The death and life of great American cities. Random House Vintage Books, New York
Kruft H-W (1996) A history of architectural theory from Vitruvius to the present. Princeton Architectural Press, New York
Lagueux M (2004) Ethics versus aesthetics in architecture. Philos Forum 35(2):117–133
Lampugnani VM (2006) The city of tolerant normality. In: Baird G, Graafland AD (eds) Cross-over: architecture, urbanism, technology. 010 Publishers, Rotterdam, pp 294–311
Latour B (2004) Why has critique run out of steam? From matters of fact to matters of concern. Crit Inq 30(2):225–248
Latour B, Yaneva A (2008) Give me a gun and i will make all buildings move: an ANTs view of architecture. In: Geiser R (ed) Explorations in architecture: teaching, design, research. Birkhäuser, Basel, pp 80–89
Le Corbusier (1986) Towards a new architecture (trans: Etchells F). Dover Books, New York [orig. Vers Une Architecture, 1923]

Loos (1908) Ornament and crime. Reprinted in: Conrads U (1970) Programs and manifestoes on 20th-century architecture (trans: Bullock M). MIT Press, Cambridge, MA [orig. Programme und Manifeste zur Architektur des 20. Jahrhunderts (1964)]
Lupton E (1996) Mechanical brides: women and machines from home to office. Princeton Architectural Press, New York
Mattsson H, Wallenstein S-O (eds) (2010) Swedish modernism: architecture, consumption and the welfare state. Black Dog Publishing, London
Moore SA, Wilson BB (2013) Questioning architectural judgment: the problem of codes in the United States. Routledge, New York
Nesbitt K (1996) Theorizing a New Agenda for architecture. An anthology of architectural theory 1965–1995. Princeton Architectural Press, New York
Norberg-Schulz C (1979) Genius loci: towards a phenomenology of architecture. Rizzoli, New York
Pallasmaa J (1996) The eyes of the skin: architecture and the senses. Academy Editions, London
Pevsner N (1949) Pioneers of modern design from William Morris to Walter Gropius. Museum of Modern Art, New York. [orig. Pioneers of the Modern Movement, 1936]
Picon A (2013) Ornament: the politics of architecture and subjectivity, AD primer. Wiley, London
Ruskin J (1849) The seven lamps of architecture. Wiley, New York
Schrijver LS (2013) Architecture as an object of research: incorporating ethical questions in design thinking. In: Basta C, Moroni S (eds) Ethics, design and planning of the built environment, vol 12, Urban and landscape perspectives. Springer, Dordrecht
Semper (1860) Der Stil in der technischen und tektonischen Kunsten; oder, Praktische Aesthetik: Ein Handbuch fur Techniker, Kunstler und Kunstfreunde. Verlag für Kunst und Wissenschaft, Frankfurt (vol 1), F. Bruckmann, München (vol 2, 1863)
Semper (2005) Style in the technical and tectonic arts (trans: Mallgrave H). Getty Publications, Los Angeles
Sennett R (2008) The craftsman. Allen Lane Books, New York
Simmel G (1903) The metropolis and mental life. Reprinted in: Bridge G, Watson S, (eds) (2002) The blackwell city reader. Blackwell, Oxford
Spuybroek (2011) The sympathy of things: John Ruskin and the ecology of design. NAi Publishers, Rotterdam
Ungers OM (1980) Architecture's right to an autonomous language. In: Portoghesi P et al. (ed) The presence of the past: First international exhibition of architecture, Venice Biennale, exhibition catalogue, London, Academy Editions, 1980
Ungers OM and L (1972) Kommunen in der Neuen Welt 1740–1972. Kiepenheuer & Witsch, Köln
van den Hoven J (2013) Architecture and value-sensitive design. In: Basta C, Moroni S (eds) Ethics, design and planning of the built environment, vol 12, Urban and landscape perspectives. Springer, Dordrecht
van Eck C (2009) Figuration, tectonics and animism in Semper's Der Stil. J Archit 14(3):325–337
van Gerrewey C, Patteeuw V, Teerds H (eds) (2013) What is good architecture? OASE, p 90, Rotterdam, NAi010 Publishers
Viollet-le-Duc EE (1875) Discourses on architecture (trans: Van Brunt H). J.R. Osgood, Boston [orig. Entretiens sur l'Architecture, vol 1, 1863, vol 2, 1872]
Vitruvius (1960) The ten books on architecture (trans: Morgan MH). Dover Publications, New York [orig. transl. Harvard University Press, 1914; manuscript Vitruvius typically dated 32–27 BC]
Watkin D (1977) Morality and architecture. Clarendon, Oxford
Watkin D (2001) Morality and architecture revisited. University of Chicago Press, Chicago
Wigley M (1988) Deconstructivist architecture. Museum of Modern Art, New York
Wigley M (2001) White walls, designer dresses: the fashioning of modern architecture. MIT Press, Cambridge, MA

# Design for Values in the Armed Forces: Nonlethal Weapons and Military Robots

Lambèr Royakkers and Sjef Orbons

## Contents



**Abstract**

Since the end of the Cold War, Western military forces became frequently involved in missions to stabilize conflicts around the world. In those conflicts, the military forces increasingly found themselves operating among the people. The emerging need in military interventions to prevent casualties translated into a range of value-driven military technological developments, such as military robots and nonlethal weapons (NLW). NLWs are characterized by a certain

This research is part of the FP7 research project "Suicide Bomber Counteraction and Prevention" (SUBCOP), which is supported by the European Union under project reference 312375.

L. Royakkers (✉)
Technical University of Eindhoven, Eindhoven, The Netherlands
e-mail: l.m.m.royakkers@tue.nl

S. Orbons
Nederlandse Defensie Academie, The Hague, The Netherlands
e-mail: sjeforbons@hotmail.com



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_28

technological and operational design "window" of permissible physiological effect, defined at each end by values: one value is a controlled physiological impact to enforce compliance by targeted individuals and the other value is the prevention of inflicting serious harm of fatality. Robot drones, mine detectors, and sensing devices are employed on the battlefield but are operated at a safe distance by humans. Their deployment serves to decrease casualties and traumatic stress among own military personnel and seeks to enhance efficiency and tactical and operational superiority.

This chapter points out that societal and political implications of designing for values in the military domain are governed by a fundamentally different scheme than is the case in the civil domain. The practical cases examined illustrate how values incorporated in military concept and system designs are exposed to counteraction and annihilation when deployed in real-world operational missions.

**Keywords**

Nonlethal weapons • Military robots • Military ethics • Designing for values • Value sensitive design

## Introduction

The end of the Cold War marked the beginning of a new era in the international security arena. In the decades before, the East-West confrontation, with its epicenter in Central Europe, had dominated military thinking and planning, and the military balance was predominantly built on the mutual large-scale destruction potential of military structures and arsenals.

The emergence of enabling technologies during the 1980s, in particular in the area of information, communication, and computing, has introduced military precision strike capabilities, implemented as precision-guided munitions and missiles (PGMs), and capable of autonomously finding and striking targets at long range. Technological advances in the area of information and communication also enabled the introduction of the so-called network-enabled capabilities (NEC). In NEC, military command and control systems are integrated with a variety of sensor platforms collecting military data and with PGMs. Such "system of systems" provided for a dramatic increase both in effectiveness and in efficiency in warfare. Rather than introducing fundamentally new technologies, military innovation focused on *system* technology, to optimally exploit and combine the potential of emerging civil technologies in novel military system concepts. Fielding such military systems concepts also entailed an increase in automation of military tasks and functionalities. This led to a shift in responsibilities in military decision-making processes.

The technological advances in the military domain were first applied at large scale during the First Gulf War in 1990. The operations in Iraq not only demonstrated the effectiveness of long-range precision attacks, but at the same time these new military-technological capabilities intrinsically entailed the value of drastically

reducing the number of military casualties on the side of the intervention forces. Thus, the design and fielding of a new family of long-range PGMs had, alongside the significant increase in military effectiveness, served a key value: the protection of the life of troops deployed in expeditionary military missions. [1]

Traditionally, in the military debate the above innovations are often referred to as the Revolution in Military Affairs (RMA): the emerging technologies provided for more precision and discrimination in the application of military firepower causing less collateral damage and fewer own causalities (see, e.g., Freedman 1998; Latham 1999). They support the two most important values on which the *jus in bello* – which involves the legal standards that apply during the fight and monitors the way in which war is waged – is based: discrimination and proportionality. The principle of proportionality states that the applied force must be proportional to legitimate military goals. Civilian casualties are acceptable as collateral damage if it is proportionate to a legitimate military advantage. The principle of discrimination states that in target selection a distinction must be made between combatants and noncombatants and between civilian objects, e.g., hospitals and churches and military objectives. Soldiers who are injured or have surrendered should cease to be targets. These two values have been interpreted and materialized in, for instance, international humanitarian law and in international treaties banning, regulating, or limiting the possession and use of particular forms of weaponry. The principle of discrimination forbids the intentional killing of "the innocent," and the underlying idea is that civilians should not be made to suffer in war; overall, this is a rights-based principle. The principle of proportionality is a consequentialist one and requires that enemy combatants should not be subjected to unnecessary suffering and superfluous injury, that it is unjust to inflict greater harm than that which is unavoidable in order to achieve legitimate military objectives, and that a mission is permitted only if the excepted military gain outweighs the expected number of unintended civilian casualties.

Since the end of the Cold War, Western military forces became frequently involved in missions to stabilize conflicts around the world. In those conflicts, the military forces increasingly found themselves operating among the people, often in built-up areas, with opposing militant forces exploiting this environment as cover. Former General Sir Rupert Smith introduced a new paradigm that contemporary military forces were now facing "the war among the people" (Smith 2006) or asymmetric warfare. In this complex environment, with blurring distinction lines between combatants and noncombatants, the casualty aversion norm for own military personnel was soon extended to include the protection of the lives of the civilian population in conflict areas as well. Some scholars predicted and claimed that from now on warfare would be conducted "humane" and, ultimately, "bloodless" (Coker 2001; Toffler and Toffler 1994). A new value was born, sharply contrasting against the armed forces' core business of killing and destroying the safeguarding of citizens during armed operations.

[1] In hindsight, however, the follow-up in Iraq between 2003 and 2011 of the First Gulf War was much more lethal, as the nature of the conflict had become irregular and asymmetrical, thus marginalizing the role of PGMs.

The emerging need in military interventions to prevent innocent casualties among the local population translated into a range of value-driven military technological developments, such as military robots and nonlethal weapons (NLW). Whereas proponents of the concept exclaimed high expectations of this new category of military capabilities, empirical analysis, both into military robots and NLW deployments in recent operations, reveals that the operational effect incorporating the intended value is flawed and, in some cases, even reversed. Other than is often the case in the cooperative and socially benign civil domain, where interests and values are the subject of constructive dialogue, in the military conflict domain, different actors have, almost by definition, sharply different and competing interests. These actors are noncooperative and even hostile toward the other side's operational objectives and focus on de-optimizing the opponent's capabilities, including those incorporating self-imposed value-based military effect characteristics. Insurgents, for instance, attempt to create conditions in such a way that the value-based purpose of preventing innocent civilian casualties is denied by bringing innocent civilians close to or inside the legitimate military target, without the operator of an armed robot or the robot itself able to detect this. Similarly, opponents to security forces that apply NLWs against them may use countermeasures to neutralize the NLW effect, which in turn may bring security forces to use the NLWs beyond its safety margins, thus risking civilian casualties. Hence, in the military domain, the functional intent of value sensitive design (VSD) of NLWs and robots is undermined by noncooperativeness and counteraction. The implication of such VSD denial is the loss of credibility of weapon user presenting themselves as protector of the innocent population. The VSD preemptive mechanisms reflect the essence of military conflict, which is an armed clash of interests.

This chapter will point out that societal and political implications of VSDs in the military domain are governed by a fundamentally different scheme than is the case in the civil domain. The practical cases examined here illustrate how values incorporated in military concept and system designs are exposed to counteraction and annihilation when deployed in real-world operational missions. In section "Non-lethal Technologies and Weapon Concepts" we will discuss the nonlethal weapons and in section "Military Robots" military robots. We will end with some conclusions.

## Nonlethal Technologies and Weapon Concepts

Although the notion of nonlethal weapon (NLW) was already coined in the first half of the twentieth century, in the military domain, it made a rebirth in the early 1990s.[2]

[2]The term *nonlethal weapon* already appeared in writings on colonial policing during the 1930s (Gwynn 1934, pp. 32–33).

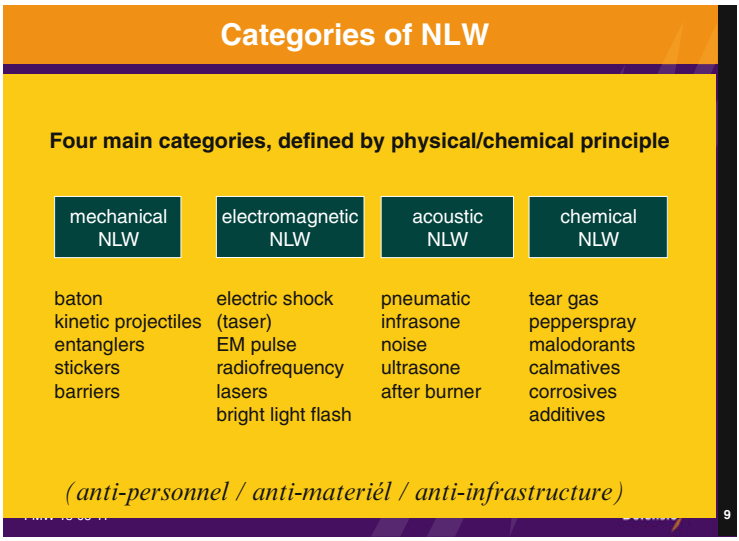



**Fig. 1** Taxonomy and examples of NLWs

NLWs are designed and deployed with the purpose to enforce change and correction of human behavior, in order to achieve people's compliance with orders or directions, without causing (innocent) casualties or serious and permanent harm to people.

Several definitions of NLWs exist. A broadly accepted one comes from NATO, which defines NLWs as:

> Weapons which are explicitly designed and developed to incapacitate or repel personnel, with a low probability of fatality or permanent injury, or to disable equipment, with minimal undesired damage or impact on the environment.[3]

## Description of Nonlethal Weapons

Several dozens of types of NLW are in use or under development, either designed for use against people or against material and infrastructure (Fig. 1). The technologies and associated types of physiological effects of NLWs are wide ranging. They are arranged into four categories, namely, kinetic energy concepts, electromagnetic effectors, acoustic energy concepts, and chemical and biological effectors. Various types from all four categories are already in use for many years and occasionally even for decades, with the police and law enforcement organizations being the forerunners in fielding NLWs before military organizations did. Long-standing NLWs are:

[3]NATO: NATO Policy on Non-Lethal Weapons, NATO, Brussels (13 Oct 1999.

- *Kinetic energy* NLWs such as the baton round, a cylindrically shaped PVC projectile designed to be fired against an individual to cause pain and blunt trauma, or the bean bag, a small sack filled with pellets launched from a small caliber projectile to cause a similar effect when striking the human body
- *Chemical* NLWs such as tear gas (CS) that has an irritating effect on the eyes, skin, and airways
- *Electromagnetic energy* NLWs like the Taser that causes a muscular incapacitation effect by electrical current
- *Acoustic* NLWs such as the fighter aircraft afterburner that can be used as an acoustic weapon and the acoustic hailing device consisting of an array of loud speakers to produce a focused high-energy noise beam or to use for messaging.

In addition, a considerable number of NLWs are under development, in engineering or testing phase, or redesigned. The Active Denial System, laser warning devices, and extended effect flash/bang grenade are a few examples. This chapter addresses antipersonnel NLWs only.

NLWs intended for use against people are tailor-made to inflict pain or other forms of physical discomfort. The intensity of the unpleasant sensation should pass a certain minimal threshold to accomplish a particular behavioral effect, which at the same time should remain below a certain maximum for safety reasons, to ensure that the physiological effect is nonlethal. While these requirements are incorporated in the technological design of the NLW, they also rely on the methods, procedures, and tactical guidelines for military personnel on when and how to operate an NLW. Each NLW can be characterized by a certain technological and operational design "window" of permissible physiological effect, defined at each end by values: one value is a controlled physiological impact to enforce compliance by targeted individuals, and the other value is the prevention of inflicting serious harm or fatality.

Ever since their inception, NLWs have been the subject of intensive debate. To some analysts and commentators, their emergence raised high expectations in reducing the number of civilian casualties in military missions, symbolized by some proponents portraying NLWs as "weapons of mass protection" (Morris 1992). Such optimism coincides with the responsibility felt in Western states to humanize war and to comply with the associated imperative of casualty aversion that is amplified by media presence in conflict zones (Coker 2001, p. 18). Others, such as McNab and Scott (2009), add that in an increasingly complex operational environment, NLWs may reduce the level of violence (US) that forces incur as well as experience in asymmetric warfare. Innovative NLW concepts have also been claimed to be promising for military tasks, due to their potentially broad applicability and to the sheer novelty of the technologies applied (Gompert et al. 2009, pp. 95–110).

Such claims are disputed by skeptics, who stress the unreliability of NLWs on the basis of accounts of incidents in which the application of NLWs led to severe harm or even fatal injury to individuals. In most of the cases NLWs were mostly used by the police. Such opposing views are reinforced by reports from human

rights organizations stressing the excessive use, or abuse, of such devices by law enforcement agencies (Amnesty International 2004). In addition, the use of NLWs against civilians by military forces abroad has been disputed on moral grounds, as it would potentially violate the principle of noncombatant immunity that considers noncombatants as complete "outsiders" in armed conflict and should therefore not be harmed (Mayer 2007).

A key question underlying this debate is whether existing and novel NLWs meet their promises under real-world conditions. Where NLWs are claimed to help manage violence in the complexity of today's operational environment, in reality this complexity may also backfire against NLW performance.

Experiences with NLW in real-world operations reveal that in many events dynamics are at work that tend to put the design window of permissible action under pressure. The reason for this is that target individuals may decide to develop countermeasures to reduce or neutralize the NLW physiological effect. In addition, the user, security forces, who are tasked to control a riot or public disorder, may be inclined to use the NLW at their disposal excessively and beyond the prescribed mode of permissible employment, in an effort to achieve the desired effect of compliance by the targeted individuals.

Mitigation of the NLW performance, induced by the dynamics at play during a real event, has the potential to transform the dual positive and benign value as envisioned with NLW design into an instrument for abuse and repression. The underlying premise in the design as de-escalating violent confrontations proves illusive under the presence of overriding factors in the operational context that negate the NLWs' original rationale and value.

Hereafter, two NLW system concepts will be more closely investigated, in order to assess their reliability in meeting the promise in terms of operational effectiveness, own force protection, and nonlethality as intrinsic design values. Challenges will be discussed that degenerate the intended value sensitivity of NLW design. The first system concept is a kinetic energy projectile, called the baton round (BR), a classical NLW, in use with police and military forces worldwide since several decades. The second is a millimeter-wave electromagnetic energy weapon concept, the Active Denial System (ADS), a US-developed concept, currently in the prototype testing and evaluation phase.

## Two Examples of Main NLW Technologies

### Innovating Classical Nonlethal Technologies: The Baton Round

BRs, also called "plastic bullets," are blunt impact weapons launched against individuals. BRs are cylindrically shaped, have diameters between 30 and 40 mm, are between 10 and 15 cm long, and have a rounded impact face. The purpose of the BR is to induce pain, irritation, and minimal injury, in order to dissuade or prevent a violent or potentially violent person from pursuing the intended course of action. The physiological effect depends on the area where the projectile strikes the human body (Vilke and Chan 2007, p. 342). The intended

effect on the target individual resembles the punch of a boxer. Ideally, the BR strikes the abdomen, while hits on the extremities, in particular the legs, are also effective. The delivery system for BRs is usually a handheld baton gun.

The projectile's velocity and ballistic stability are key factors for aiming accuracy. Launching velocities are around 80 m/s. Ballistic stability can be enhanced with spin stabilization of the projectile. Required accuracy on the target is usually defined as a probability that the projectile strikes in an area of 20 cm wide and 60 cm high. In 2004 a report of a UK program for an improved BR set this probability at 85 % for a minimal distance of 25 m and desirably up to 40 m. Desirable accuracy on a target should be 20 cm wide and only 40 cm high, with the aiming center on the abdominal part of the body (UK Steering Group 2004, pp. 11–18).

BRs have been deployed in many law enforcement and military forces around the world. In Northern Ireland, during the Troubles lasting from 1969 until 1998, more than a hundred thousand BRs have been fired by the British Army and the Royal Ulster Constabulary. During these three decades and thereafter, several technological innovations have been implemented to improve the performance and reliability of the BR. The round was introduced in Northern Ireland in the mid-1970s. Until then rubber bullets were deployed, with lower ballistic accuracy standards. A medical report on injuries caused by rubber bullets states that due to firing the round from a tear gas (CS) canister launcher, the tumbling of the projectile in flight and poor aerodynamic shape, it was difficult to hit at 18 m a target with a 2 m diameter (Millar et al. 1975, p. 480). Early versions of the BR deployed in Northern Ireland also had relatively low performance standards, which were gradually improved through successive innovative designs (Burrows 2002, pp. 105–107).

The BR's potential for raising its performance is facing difficult challenges. Effectiveness at ranges above 50 m is poor. Kinetic energy drops significantly at longer ranges, at 25 m to about 75 % of the level at a range of 10 m. At longer engagement ranges, the flight trajectory of the round is more curved, reducing aiming accuracy (Arnesen and Rahimi 2007). Shorter firing ranges enhance accuracy, but deliver a heavier impact on the target, thereby increasing the potential of injury.

Efforts to further improve baton rounds are ongoing. In the UK, for instance, a new projectile has been developed that should be safer and more reliable than its predecessors. This projectile, the Attenuating Energy Projectile (AEP), was introduced in the UK in 2005, to replace a more hazardous predecessor (UK Steering Group 2006, pp. 15–20).

BRs are generally employed to meet two values: to provide security forces with a capability to outrange missile throwing rioters and other violent actors, thus enabling the use of armed force without having to resort to lethal fire for self-defense and at the same time forcing revolting individuals to stop their violent behavior.

Over time, seasoned rioters managed to develop countermeasures to negate the BR effect, such as makeshift body protection and evading tactics. Security forces, facing the declining effectiveness of the NLW, were in many situations tempted to use the weapon beyond its safety margins to acquire effect, while putting targets at risk of

serious harm, thus compromising the design value of nonlethality. Contextual issues, interfering with the attitude and behavior of security personnel operating the BRs, have led to prohibitive use and reckless abuse of the NLW. As a result, in Northern Ireland, inappropriate use has importantly contributed to the death of 17 civilians by BRs and permanent harm to many hundred victims (Weir 1983, p. 83).

Any conceivable effort to cope with countermeasures against BRs and to maintain the window of permissible use of the BR most likely requires the introduction of smartness in the technological design of the BR. Smart BRs should be capable of autonomously identifying the shape of the engaged target, to fine-tune the kinetic impact energy and to "decide" where precisely to strike the body. The feasibility of such value restoring innovation hinges on the affordability of the relatively large numbers of BRs usually required.

### Introducing a Novel Nonlethal Technology: The Active Denial System

The millimeter-wave (MMW) directed energy technology, called the Active Denial System (ADS), is a unique NLW concept developed in the USA. The weapon effect can attain ranges of many hundreds of meters to engage human targets. Its MMW beam is invisible and no traces when properly employed. At the same time, the effect mechanism is entirely new, and, other than with most "first-generation" NLWs, there is no precedent available with law enforcement agencies.

The ADS delivers a totally different type of effect. There are no empirical data available to which a military planner, commander, or operator can refer, other than the many tests and experiments that have been conducted to map human bio effects (Murphy et al. 2003).

The development of the MMW technology for the ADS started already in the early 1990s. It is based on 94 GHz radiation emitter technology, and the radiation beam interacts with the human body such that it penetrates the skin to a depth of less than 1/64th of an inch or less than half a millimeter. The beam shape can be adjusted to engage between one and four target individuals at a time. A targeted person will experience the effect as a sharp burning pain on the skin but no actual burning, which he immediately wants to escape by jumping aside or running away. This pain effect is universal, as it is independent from the size or physical condition of the target individual.

The system concept's particular value is the exceptional long range at which it can deliver its effect. Much more than with other NLWs, the own troops can stay out of harm's way. The second value entails that it forces target persons to comply with the security forces' directions, due to the intolerable pain it causes. Hence, if the system is capable to deliver radiation energy intensities with sufficient accuracy at the target over hundreds of meters, it offers a promising perspective for being capable of serving both the envisioned value of the force's self-protection and a reliable nonlethal impact on target individuals.

In reality, circumstances may be such that effect control is only marginally achievable. One important limiting factor is that the radiation energy is strongly attenuated by water; even under high ambient humidity the transmission gain will drop substantially. Rain, in particular heavy rain, will strongly reduce the radiation

energy level arriving at the target. This degradation is more significant with increasing range. Similarly, wet cloths (either self-wetted or by rain) will reduce the effect felt on the body. The complicating factor is the high uncertainty about what the residual energy arriving at the skin will actually be. This may vary by orders of magnitude, depending on the specific attenuation effect in the specific situation. The fact that this cannot be measured while the ADS is in action is problematic, and rough estimates or trial and error are not acceptable options. This would resemble Russian roulette.

The sheer novelty of the design leaves questions unanswered like "how does it work?" and "does it work as it should?" Today, most military personnel have only been trained on the use of kinetic force and when and how to use it. They can hardly grasp the practical utility of ADS in the face of the many uncertainties surrounding such a revolutionary concept. When the system does not facilitate automatic beam energy modulation, will the operator be susceptible to human error when having to tune the MMW transmitter manually under circumstances of incomplete information collection?

## Central Moral Values and Value Issues of NLWs

Both the BR and ADS discussions demonstrate that nonlethality as the defining value of the technology and design of both NLWs is challenged when the systems are deployed to feature in real-world operational events they are intended for. Experiences and assessments of the two NLWs have identified a range of factors and phenomena that, either naturally or human driven, narrow down the weapons' design window of nonlethal performance and effect. Many of those factors are pertinent to the so-called fog of war,[4] implying that much of what military forces encounter in real-world operations is unforeseeable, hence renders their capability of controlling the scenario illusive, and tends to counteract their intended approach and the accomplishment of their mission (Orbons and Royakkers 2014).

This section focuses on key value issues related to NLW design and military implementation of NLWs and claims on values of NLWs with respect to their political purpose.

### Performance of NLWs in a Military Context

A range of conditions shapes a non-cooperative environment for the user of NLWs, denying the efficacy in projecting the value of nonlethality that is embedded in NLW designs. In essence, in military operations a conflict of values is at work, which puts the military forces' responsibility for self-protection in many scenarios at odds with the requirement imposed on the forces to prevent innocent casualties.

[4]The expression of the "fog of war" was coined early in the nineteenth century by the Prussian General Carl Von Clausewitz [1831] (1984), in his famous work "Vom Kriege" ("On War"). Its relevance for NLWs has been addressed in Orbons (2010).

Hence, the extent to which the value of nonlethality embedded in NLWs is brought to bear is to a far extent in the hands of the military operator, rather than an NLW system attribution. In addition, "smart" targets have ways to overcome the intended effect of NLWs, as has been pointed out by Allgood (2009) and Hussey and Berry (2008), in the case of riots in a detainee camp under command of US forces.

While degenerated use of NLWs may result from mechanisms emerging from stress among the military forces applying the NLWs, situations have occurred in which NLWs were intentionally used in a non-regular or degenerated use. Such incidents took place in Northern Ireland, where civilians have been killed by BRs that where aimed at vulnerable parts of the body or fired from very short distances (Pat Finucane Centre 1996). Nonpermissive uses of NLWs, in particular kinetic NLWs, have been found in Iraq detention centers as well. It is difficult to determine to what extent user forces intentionally apply NLWs irregularly or not. The operational context shapes a gray zone in which the imperative of self-defense can hardly be distinguished from unnecessary and excessive harm. NLWs have been used as instruments of punishment and retribution for negative outcomes of previous events or confrontations to the user forces, as the Northern Ireland case demonstrates. A similar observation has been made in 2011, in the aftermath of the Arab Spring in Egypt.[5] Lack of discipline and training, and insufficiently restrictive instructions and Rules of Engagements for user forces, can contribute to the probability of such wrong uses to occur. It is an inherent problematic and risk of many types of NLW weapons and technologies that carry the potential for harmful and even lethal use, when specified safety margins are ignored.

Can innovative concept solutions in value design overcome the shortfalls of current NLWs? One important mechanism that underlies the "us-or-them" dilemma in asymmetric conflicts is the closing-in of the military forces with the civilian population: Rupert Smith's "war among the people" in its truest sense. Obviously, design that enables a larger standoff between force employing NLWs and the civilian population would serve to diminish the us-or-them dilemma. The ADS is actually the concept standing out as the champion technology and designed to support that aim. As we have seen, however, scenarios and circumstances are conceivable that annihilate the accomplishment of the value-based effect.

### Political Level Significance of NLWs

At the political level, mission accomplishment strategies call for instruments at the tactical level compatible with the spirit and objectives of the military mission. It means the balanced use of force, with a measured application of armed force. In situations where civilians are involved, compliance should be accomplished without causing harm. NLWs are considered and assigned as appropriate instruments for that task: they are expected to enable humane military operations and performance, in

[5]In November 2011, protestors in Cairo were killed as a consequence of asphyxiation by particular types of tear gas and others blinded or otherwise injured by rubber bullets intentionally fired at the head and neck (Human Rights Watch 2012).

support of the hearts and minds strategy. The implementation and purpose of NLWs is publicly announced: intentions and expectations are declared explicitly.

As pointed out by Orbons (2012), in real-world situations, NLW deployment is fraught with problems related to the operational context; consequently the level of control over NLW effects is much less than what is militarily and politically desired. Most soldiers are far from perfect in dealing with the dynamics and uncertainty on the ground. Moreover, the political rationale for NLW deployment is counteracted and undermined by opponents who force the military user into the lethal part of the spectrum of violence, thus annihilating the nonlethal intent.

But progress in hearts and minds efforts at the tactical level, conversely and ironically, may under circumstances also be affected by trends and events at politico-strategic level. If these trends have the effect of antagonizing user forces and target populations, the ensuing operational context will frustrate the outcome of NLW deployment as politically intended and expected. Hence, the political level rationale of NLW deployment becomes annihilated at the tactical level if particular developments at the political level meet disapproval and trigger agitation on the ground. Obviously, with regard to the nonlethality incentive, a dialectic is at work between the political and tactical level. This dialectic is fueled when operational context mechanisms as friction and confusion produce fatal errors and further amplified by the media connectivity between the tactical and political level. The tragedy is that the media are inclined to report only the mishaps (innocent casualties, despite or even caused by NLWs), while refraining from spreading good news about NLWs performing "normal" as expected and announced: good news is "bad" news.

Hence, if NLWs perform badly, their deployment backfires at the political level. If, however, the tactical/political dialectic link is weakened or cut through flaws in reporting along the chain of command or in public information, chances of optimal NLW use increase. The flip side of this condition is the growing risk of abuse due to the tactical isolation of physical engagements, as accountability mechanisms would be dysfunctional and only have a delayed political impact at best. In the latter case, abuse will surface sooner or later and will give NLWs a bad reputation after all.

In coping with the dialectic, which in essence is described by Rupert Smith's (2006) "war amongst the people" paradigm, some planners and developers search for nonlethal technological options to physically disengage the user from the target. The ADS, with its long-range and semi-area denial capability, is the ultimate material expression of this quest. However, the technology fix approach ignores that disengaging the user force from the target population is at odds with the hearts and minds approach. This reflects another dialectic, namely, that between (community) policing and military operations.

## Military Robots

In the last two decades, we have entered the era of remote-controlled military technology: robot drones, mine detectors, and sensing devices are employed on the battlefield but are controlled at a safe distance by humans. Its aim is to decrease the

number of soldiers killed on the battlefield, to gain more efficiency and tactical and operational superiority and to reduce emotional and traumatic stress among soldiers (Veruggio and Operto 2008).

All over the world military robots are currently being developed, and thousands of military robots are already deployed during military operations. According to Peter Singer this development forms the new "revolution in military affairs" (Singer 2009). The US *Future Combat Systems*, a US $200 billion plus program for future weapons and communications systems – commissioned by the Pentagon – has a major impact. Military robots are a focal point in this program. Besides a technology push, a demand pull has been added for the development of military robots. The call of US society to reduce the number of military casualties has contributed to a huge boost of alternatives in robotics developments in the USA.[6] A few years ago, the number of US soldiers killed in action rose to a high level because of insurgents operating in Iraq and Afghanistan using their popular homemade and deadly weapon: the improvised explosive device (IED) or the roadside bomb. 40 % of those killed American soldiers died because of these IEDs (Iraq Coalition Casualty Count 2008). During the invasion of Iraq in 2003, no use was being made of robots, as conventional weapons were thought to yield enough "shock and awe." However, thousands of American soldiers and Iraqi civilians killed reduced popular support for the invasion and made the deployment of military robots desirable. By the end of 2008, there were 12,000 ground robots operating in Iraq, mostly used to defuse roadside bombs, and 7,000 reconnaissance planes or drones were deployed (Singer 2009).

## Description of Military Robots

We will define military robots as reusable unmanned systems for military purposes with any level of autonomy. It may involve both unmanned systems that are self-propelled, i.e., mobile robots, as well as static systems that perform tasks, as in immobile robots. An example of an immobile robot is the operational *Goalkeeper* on board of Dutch frigates. This is a computerized air defense system with infrared detection of enemy missiles which autonomously detects approaching missiles, calculates the path they follow, and then aims the weapon, being a rapid-fire gun, in order to neutralize the approaching danger.

Military mobile robots are commonly divided into ground vehicles, water surface and underwater vehicles, and aerial vehicles. The most famous unmanned ground vehicle, which has been developed by Foster-Miller, is SWORDS (*Special Weapons Observation Reconnaissance Detection System* – see Fig. 4). It originated from TALON, a robot equipped with cameras, a gripper arm, communication,

[6]Former Senator John Glenn once coined the term "Dover Test": whether the public still supports a war is measured by responses to returning body bags. He called it the "Dover Test" as the coffins of killed American soldiers came in from abroad at the air base in Dover, Delaware.

distraction devices, and various sensors – thus a device especially designed for unmanned reconnaissance and clearing roadside bombs. SWORDS is equipped with machine guns and a remotely controlled, tele-led armed robotic system. After some years of research, a number of SWORDS have been deployed since 2007 on patrols in Iraq. These SWORDS mainly perform reconnaissance missions, street patrols, and other missions with an increased risk. The successor of the SWORDS already exists in the MAARS, Modular Advanced Armed Robotic System. This robot can be equipped with heavier machine guns, has a larger payload, and has nearly twice the speed at about 11 km/h. The entire system weighs about 160 lb.

Unmanned submarines equipped with torpedoes are currently being developed. Existing unmanned mini-submarines can autonomously explore the seabed with sensitive listening devices, detect ships and mines, and destroy those mines with an explosive charge. Unmanned vessels such as the 9 meter *Protector* or the nearly 2-m long *Silver Marlin* are equipped with sensors, a satellite connection, and light armament that can take over patrols from small warships.

An example of unmanned aerial vehicles is the micro air vehicles, unmanned reconnaissance helicopters. These are remote-controlled propeller planes as small as a model airplane with a weight of about 20 g to a few 100 g and equipped with powerful regular or infrared cameras for autonomous observation tasks. The camera images are so sharp that persons placing parcel bombs or roadside bombs can be detected and monitored, alerting local forces to act. Also, these aircrafts can search targets and communicate the position for conventional bombing. At the end of 2001, the US deployed about 10 unmanned reconnaissance aircraft in Afghanistan, but in 2008 these numbers had already grown to more than 7,000 (Singer 2009). Besides these small aircrafts, there is the reconnaissance Global Hawk with a wingspan of nearly 40 m. This unit can eavesdrop on mobile phone calls, even if they are encrypted, and could provide real-life images from an altitude of some kilometers, spotting a car on the road, not quite making out the license plate of the car, but surely the car type and how many people are moving around it.

Presently, more than 20,000 military robots are active in the US military. Most of these robots are unarmed and are mainly used for clearing improvised explosive devices and reconnaissance; however, over the last years the deployment of armed military robots is on the increase. In this chapter we will focus on the unmanned combat aerial vehicles.

## Unmanned Combat Aerial Vehicles (UCAVs)

One of the most widely used unmanned combat aerial vehicles (UCAVs) is the *Predator*. This unmanned airplane which can remain airborne for 24 h is currently employed extensively in Afghanistan. The *Predator* drones can fire hellfire missiles and are flown by pilots located at a military base in the Nevada Desert, thousands of miles away from the battlefield. On the top of this its successor, the *Reaper*, which may phase out the F-16, has already been spotted in Afghanistan in 2008.

This machine with a wingspan of 20 m can carry 5,000 lb of explosive devices, Hellfire missiles, or laser-directed bombs and uses day-and-night cameras to navigate through a sheet of clouds. This unmanned combat aerial vehicle is operated by two pilots located at a ground control station behind a computer at a safe distance from the war zone.

As if the tactical advantages brought by this technology were not enough, we now face the prospect of genuinely autonomous robot vehicles, those that involve "artificial intelligence" and hence do not need human operators. This shift is also stimulated by the National Research Council (2005): "The Navy and Marine Corps should aggressively exploit the considerable warfighting benefits offered by autonomous vehicles." The United States Air Force (2009), for example, expects the deployment of autonomous UCAVs with "a fully autonomous capability" between 2,025 and 2,047. Though it is unclear what degree of autonomy these UCAVs will have, "the eventual deployment of systems with ever increasing autonomy is inevitable" (Arkin 2009). The deployment of genuinely autonomous armed robots in battle, capable of making independent decisions as to the application of lethal force without human control, and often without any direct human oversight at all, would not only constitute a genuine military revolution, but also a moral one (Kaag and Kaufman 2009).

Given the distinction between, on the one hand, UCAVs today, in which – to differing degrees – human operators remain in the loop, and, on the other, the future of military robotics which promises autonomous UCAVs capable of ethical decision-making, we will try to separate our analysis along these lines.

### Tele-operated UCAVs

In the relevant literature the role of the human operator is often underplayed. The importance of having an element of human control incorporated in the design of UCAVs has often been stressed, for example, by the Pentagon or the British Ministry of Defence (Krishnan 2009). From a legal and ethical perspective, the value of keeping the "man-in-the-loop" is important because it is indispensable for the attribution of responsibility (cf. Singer 2009). It is not without reason that the "International Law of Armed Conflict dictates that unmanned systems cannot fire their weapons without a human operator in the loop" (Isenberg 2007). Yet, while it is certainly true that currently humans are kept "in-the-loop," it is not certain, or even likely, that this will remain so. The logic that brought unmanned systems into being leads more or less naturally to the wish to take the human out of the system altogether (Sparrow 2011, p. 121; see also Sullins 2010), and it seems almost a given that the future will hold autonomous and even learning robots.[7] We will turn to these autonomous and learning robots in the next subsection.

[7] In fact, the USA expects to operate autonomous robots in 2035 (US Department of Defense 2009), while South Korea already has autonomous robots, stationary but armed with a derivative of the FN Minimi – a light machine gun, capable of fully automatic fire – guarding the border of North Korea.

The tele-operated UCAVs connect the human operators with the war zone; they are the eyes of the tele-soldier. These semiautonomous UCAVs (they can navigate to their goal auto controlled, but the decision to fire is made by a human operator) like the Predator and the Reaper send GPS coordinates and camera images back to the operator. Based on the information projected on his computer screen, the interface, the human operator has to decide, for example, whether or not to launch a missile. With regard to the user interface design, display characteristics, interaction mechanisms, and control limitations all have a potentially huge impact on the situational awareness of the human operator and decision-making by the human operator.[8]

There is a growing concern about and interest in the ethical design of weapon systems interfaces and lethal tele-operated systems (see, e.g., Asaro 2009; Cummings 2006). In the future, his decision might be mediated by a computer-aided diagnosis of the war situation (see also Sullins 2010, p. 268), and military robots may even have ethical constraints built into their design – a so-called ethical governor, which suppresses unethical lethal behavior. For example, Arkin (2009) has done research (sponsored by the US Army) to create a mathematical decision mechanism consisting of prohibitions and obligations derived directly from the laws of war. The idea is that future military robots might give a warning if orders, according to their ethical governor, are illegal or unethical. For example, a military robot might advise a human operator not to fire because the diagnosis of the camera images tells it and the operator is about to attack noncombatants, i.e., the software of the military robot that diagnoses the war situation provides the human operator with ethical advice to support values as limiting civilian deaths and war crimes or atrocities. The software must function reliably in complex and dynamic environments, and its ethics cannot simply be a list of rules or norms as the situations that most often require ethical decision-making are exceptional cases where the standard rules or norms do not apply.

### Autonomous UCAVs

The ultimate goal of autonomous military robots, according to United States Air Force (2009), is to create a military robot capable of making independent decisions as to the application of lethal force without human control, in other words, to strive for the man-out-of-the-loop. We need to make a distinction for these autonomous robots between non-learning machines and learning machines. Learning military robots, based on neural networks, genetic algorithms, and agent architectures, are able to decide on a course of action and to act without human intervention.[9]

[8]For the impact on situational awareness, we refer to Riley et al. (2010).

[9]Although learning armed military robots appear high on the US military agenda (Sharkey 2008), the deployment of these robots is, at least under present and near-term conditions, not reasonable within the next two decades (Arkin 2009). Barring some major significant breakthrough in artificial intelligence research, situational awareness cannot be incorporated in software for lethal military robots (Gulam and Lee 2006; Fitzsimonds and Mahnken 2007; Kenyon 2006; Sharkey 2008; Sparrow 2007).

The rules by which they act are not fixed during the production process, but can be changed during the operation of the robot, by the robot itself (Matthias 2004). The problem with these robots is that there will be a class of actions where no one is capable of predicting the future behavior of these robots anymore. So, these robots would become a "black box" for difficult moral decisions, preventing any second-guessing of their decisions. The control transfers then to the robot itself. This will constitute a responsibility gap (Matthias 2004), since the value responsibility cannot be added in a design for autonomous learning military robots. It would constitute the injustice of holding people responsible for actions of robots over which they could not have any control (see also Sparrow 2007).[10]

The learning machines for military purposes seem, at least under present and near-term conditions, far from feasible. However, the development of autonomous non-learning machines with an additional ethical dimension is a newly emerging field of machine ethics. These robots, based on syntactic manipulation of linguistic symbols with the help of formal logic, are "able to calculate the best action in ethical dilemmas using ethical principles" (Anderson and Anderson 2007). It is thus assumed that it is sufficient to represent ethical theory in terms of a logical theory and to deduce the consequences of that theory. This view, analogous to the reduction of ethics to law or reflection to an algorithm, misunderstands the unique – nonreducible – nature of ethical reflection. Arkin (2009) argues that some ethical theories, such as virtue ethics, do not lend themselves well by definition to a model based on a strict ethical code. While military robotic specialists claim that the solution is simply to eliminate ethical approaches that refuse such reduction, we argue that this nonreducibility is the hallmark of ethics. While many ethical situations may be reducible, it is the ability to act ethically in situations that call for judgment that are distinctly human. Furthermore, a consequence of this approach is that ethical principles themselves will be modified to suit the needs of a technological imperative: "Technology perpetually threatens to coopt ethics. Efficient means tend to become ends in themselves by means of the "technological imperative" in which it becomes perceived as morally permissible to use a tool merely because we have it" (Kaagman and Kaufman 2009).

## Central Moral Values and Value Issues of Tele-operated UCAVs

The use of UCAVs provides us with an ambivalent picture. On the one hand the deployment of these robots has many positive effects. The most compelling arguments in favor of UCAVs are the decreasing of financial costs; reducing of the number of military casualties; added value in performing dull, dangerous and dirty tasks to solve operational problems; and effective and efficient performance of tasks. According to Strawser (2010) in certain circumstances the use of armed

[10]Schulzke (2013) has argued that it is possible to attribute responsibility to autonomous robots by addressing it within the context of the military chain of command.

military robots, for the reasons mentioned above, is not only ethically permissible, but instead even ethically mandatory under what he calls the "principle of unnecessary risk." Strawser argued that it is morally reprehensible to command a soldier into running the risk of fatal injury, if that task that could also have been carried out by a military robot. UCAVs, however, also raise all kinds of social and ethical questions that are of importance in a responsible use of these weapons. In this section we will discuss some value issues related to the decision-making process involving life and death by the human operators of tele-operated UCAVs and by autonomous UCAVs.

### Reducing Psychological Stress

The human operators, who remotely control armed military robots by computers, can be emotionally and psychologically affected by the things they see on screen. Although fighting from behind a computer is not as emotionally potent as being on the battlefield, killing from a distance is still stressful; various studies have reported physical and emotional fatigue and increased tensions in the private lives of military personnel operating the Predators in Iraq and Afghanistan (Donnelly 2005; Kaplan 2006). For example, a drone pilot may witness war crimes yet find himself in a situation in which he is helpless to prevent it, or he may see how civilians are killed by his own actions. The latter is not an entirely hypothetical situation. This problem of "residual stress" of human operators has led to proposals to diminish these tensions. In particular, the visual interface can play an important role in reducing stress; interfaces that only show abstract and indirect images of the battlefield will probably cause less stress than the more advanced real images (Singer 2009). From a technical perspective this proposal is a feasible one, since it will not be hard to digitally recode the war scene in such a way that it induces less moral discomfort with the war operator. From a moral point of view, this would mean that a soldier gets detached even further, both physically and emotionally, from his actions then is presently the case (cf. Royakkers and Van Est 2010). This detachment reduces or even eliminates the stress of human operators, but also limits reflection on their decisions leading to human operators not being fully aware of the consequences of their decisions. Instead, they are only focused on the outcome, for example, the targeting of the blips on a screen, and it has to be feared that killing might get a bit easier (see also Singer 2009, pp. 395–396; Sparrow 2009, p. 179). The last observation brings us to the important role of dehumanization, i.e., seeing people for something less than humans, in making unethical conduct more likely to occur (Bandura 1999). So, the value "reducing the psychological stress on remote operators" by dehumanization may come at odds with the value "preventing unethical conduct by remote operators."

Almost 20 % of the soldiers returning from Iraq or Afghanistan have posttraumatic stress disorder or suffer from depression (cf. Tanielian and Jaycox 2008) causing a wave of suicide, particularly among American veterans that have fought in Afghanistan or Iraq. Since remotely controlled devices can reduce stress, they could also enable more humane decision-making by soldiers. It is well known that in the heat of battle, the minds of soldiers can become clouded with fear, anger,

or vengefulness, resulting in unethical behavior or even war crimes. A 2006 survey done by the US Army Surgeon General's Office (2006) confirmed this picture. Remote-controlled robotic warfare thus might have some fundamental advantages, as it distances soldiers from direct physical contact with some of the sources of this emotional stress. To further the goal of minimizing military casualties and stress-related casualties, Arkin has proposed equipping military robots with an artificial conscience that would suppress unethical lethal behavior by adding an ethical dimension to these robots. This ethical dimension consists of prescriptive ethical codes which can govern its actions in a manner consistent with the Laws of War and Rules of Engagement. Arkin (2009) stated that "they [robot soldiers] can perform more ethically than humans are capable of," because they have no revenge motive.[11] While Arkin's statement may seem like science-fiction to most, the fact is that the deployment of military robots or unmanned semiautonomous vehicles is rapidly growing.

### Responsibility

A value issue related to what constitutes an ethical design is that the ethical governors proposed by Arkin may form a "moral buffer" between human operators and their actions, allowing them to tell themselves that the military robot took the decision. According to Cummings (2004, p. 30), "[t]hese moral buffers, in effect, allow people to ethically distance themselves from their actions and diminish a sense of accountability and responsibility." A consequence is that humans then simply show a type of behavior that was desired by the designers of the technology instead of explicitly choosing to act this way and thus over-rely on military robots (the "automation bias"). This can lead to dangerous situations since the technology is *imperfectly reliable*, so the human operator must intervene when some aspect of the technology fails (Wickens et al. 2010). The values safety and keeping the man-in-the-loop with the related value responsibility are at stake here. According to several authors (e.g., Sparrow 2007; Fieser and Dowden 2007; Sharkey 2008; 2010, and Asaro 2007), the assumption and/or allocation of responsibility is a fundamental condition of fighting a just war is that an individual person may be held responsible for civilian deaths in the course of it, and that this condition is one of the requirements of *jus in bello*. Ethical governors might blur the line between nonautonomous and autonomous UCAVs, as the decision of a human operator is not the result of deliberation, but is mainly determined or even enforced by a military robot. In other words, human operators do not have sufficient freedom to make independent decisions, which makes the attribution of responsibility difficult. The moralizing of the military robot can deprive the human operator from controlling the situation; his future role will be restricted to monitoring. The value

[11]Johnson and Axinn (2013) have countered Arkin's statement and argued that robots with no emotions do not have the attitude toward people that "healthy" humans are expected to have, and that therefore well-trained humans with healthy emotions are more desirable than autonomous robots.

"keeping the man-in-the-loop" will then be eroded and replaced by "keeping the man-on-the-loop." This can have consequences for the question of responsibility: Detert et al. (2008) have argued that people who believe that they have little personal control in certain situations – such as those who monitor, i.e., who are on-the-loop – are more likely to go along with rules, decisions, and situations even if they are unethical or have harmful effects. This would imply that it would be more difficult to hold a human operator reasonably responsible for his decisions, since it is not really the operator that takes the decisions, but a military robot (see Royakkers and Van Est 2010).

### Moral Reflexivity and "Better" Ethical Decisions

Within a military context, reflexivity is essential for ethical decision-making for fundamental reasons. In order for moral judgements to be legitimate, they must be the result of a careful process of moral reflection. This entails that the determinations made by military robotics which are based on algorithms are not forms of moral deliberation and reflection. While it is clear that military robotics are capable of processing a greater amount of information at a much faster rate than human beings (which is the reason so many members of the military community are greatly in favor of drones), this ability is distinct from the ability to critically evaluate this information and to consider it when making difficult strategic decisions. Knowledge, the process of transforming information into understanding, is a skill that only human beings are capable of. Robots lack the ability to ask themselves questions about their own choices, actions, and how their interactions affect those of their environment. As all military robots lack the ability to reflect, they lack the understanding necessary for making ethical decisions in complex and changing environments. Military robots' inability to think, to reflect, or to understand their complex situational environment has been demonstrated by the fact that they often miss important details or incorrectly interpret situations in a complex and dynamic military environment. Even the most excellent sensors can never compensate for a robot's deficient understanding of its environment (Krishnan 2009). Humans are better at discriminating targets, because they understand what a target is and when and why to target something or somebody. The lesson learned is that designing and fielding an autonomous military robot for reducing mental discomfort and reducing financial costs can be at odds with a careful process of moral reflection on a lethal decision. In our opinion the value of moral reflexivity must trump the values of reducing mental discomfort and reducing financial costs if it is at the expense of moral reflexivity. The implementation of military robots must be preceded by a careful reflection on the ethics of warfare in that warfare must be regarded as a strictly human activity, for which human beings must remain responsible and in control and that ethical decision-making can never be transferred to machines, since machines are not capable of making ethical decisions.

Ethical decision-making is thus an approach that emphasizes the importance of a process of critical understanding. This differs greatly from approaches, like that of Arkin, who examines ethics from a military robotics specialist's perspective, and therefore in terms of information. They imply that applied ethics is essentially the

application of theories to particular situations: "A machine (...) is able to calculate the best action in ethical dilemmas using ethical principles" (Anderson and Anderson 2007). This view is, however, also problematic for other reasons than mentioned above, especially in the military context. One reason is that no moral theory is universally accepted. Different theories might yield different judgments about a particular case. But even if there were one accepted theory, framework, or set of principles, it is doubtful whether it could be straightforwardly applied to all particular cases. Theory development in ethics in general does not take place independent of particular cases. Rather, theory development is an attempt to systematize judgments over particular cases and to provide a rational justification for these judgments. So if we encounter a new case, we can of course try to apply the ethical theory we have developed until then to that case, but we should also be open to the possibility that the new case might sometimes reveal a flaw in the theory we have developed so far.[12] Furthermore, the laws of armed conflict, rules of engagement, and just war tradition providing the two general ethical values for lethal decision-making, discrimination and proportionality, are open to challenges and interpretations, which depend heavily upon awareness of particular situations and may not be effectively enforceable. The rules of engagement are devised by military lawyers to suit the needs of specific operations and missions, but they often appear ambiguous or vague to military personnel who observe situations that do not always fall neatly into the distinctions made by lawyers (Asaro 2009).

However, machines can help humans to make better decisions. For example, the ethical governor introduced by Arkin can be a very useful tool if it does not remove the human from the loop. If serving as a safety mechanism preventing and warning humans from making mistakes that require an override, it might be designed in a way that it does not lead to dehumanization and loss of moral reflexivity of the human operator. Thus, designing an ethics-based systems interface should be carefully investigated (cf. Hellström 2013).

## The Essence of an Ethical Design for UCAV Systems Interface

To avoid the problems mentioned above, Cummings (2006) argues in favor of the design methods of value sensitive design, which considers the impact of various design proposals on a set of values. The relevant values in play in designing UCAV systems interfaces are reducing civilian deaths and war crimes, reducing

[12]See Van de Poel and Royakkers (2011). If ethical theories do not provide moral principles that can be straightforwardly applied to get the right answer, what then is their role, if any, in applied ethics? Their role is, first, instrumental in discovering the ethical aspects of a problem or situation. Different ethical theories stress different aspects of a situation; consequentialism, for example, draws attention to how consequences of actions may be morally relevant; deontological theories might draw attention to the moral importance of promises, rights, and obligations. And virtue ethics may remind us that certain character traits can be morally relevant. Ethical theories also suggest certain arguments or reasons that can play a role in moral judgments.

psychological stress on remote operators, meeting the criteria of discrimination and proportionality, moral reflexivity, and responsibility. The idea is that design proposals are evaluated based on this set of values. The problem with this conceptualization of what constitutes an ethical design is that it only demonstrates that a certain design is better than another design according to a given set of values, but it does not give any indication how to provide an actual ethical design for a UCAV systems interface.

Essential for an ethical design of a user interface is the understanding of ethical and psychological problems the human operators face, not just in theory but empirically. A lot of research is necessary to explore the cognitive and psychological processes the human operators employ to make ethical decisions. Furthermore, it is necessary to investigate what kind of information is useful and relevant, how information should be represented, and how much information can be dealt with, so that design systems can improve the ethical decision-making of the human operators. In other words, display characteristics, interaction mechanisms, and control limitations are of huge influence of the ethical decision process. The design system of a user interface should make transparent how, from whom, and when information was obtained and how reliable it is, enabling a human operator to make a responsible ethical decision based on moral reflection. It might be required to impose high levels of psychological stress on human operators in order to improve their ethical decision-making. According to Asaro (2009), it will be valuable to study these kinds of trade-offs through an effort "to model the moral user" by designing systems that improve the ethical decision-making of the human operators. By developing a sophisticated empirical model of human operators, we can better understand, for example, the impact of psychological stresses.

## Conclusions and Outlook

This chapter has examined the role of VSD applications in the military domain. Focusing on the innovative military system concepts of NLWs and military robots, the intended values of these systems have been assessed. Due the nature of the military context, various mechanisms are at work that tend to preempt the values these weapons are designed to yield.

Nonlethality as the defining value of NLWs comes under pressure when they are deployed in real-world events. In many situations, the operational context tends to narrow down the weapons' nonlethal design window. Key factors in this operational context reside in the domain of target behavior, in the domain of the user of the NLW, and in the physical environment. Many of those factors are pertinent to the so-called fog of war, implying that much in the operational context of NLW deployment is unforeseeable and reduces the feasibility of controlling the scenario in such a way that the NLW produces the desired effect. The "fog of war" is intimately linked to the noncooperative nature of conflict environments and to the limitations on acquiring sufficient and reliable information on factors and actors shaping this environment. In turn, this noncooperativeness generates a conflict

of values that puts the military forces' responsibility for self-protection in many scenarios at odds with the requirement to the forces to prevent innocent casualties. This "us-or-them" dilemma is to a large extent defining the extent to which the value of nonlethality embedded in NLWs is brought to bear. Rather than that nonlethality is an NLW system attribution, and it is an outcome determined by its operator, who is potentially capable to overrule the weapon's design value of nonlethality. As a consequence, situations occur, in which the deployment and use of NLW can produce an escalating cycle of violence that enhances the risk of innocent casualties, rather than reduce it; hence, it becomes counterproductive.[13]

An ethical design of a user interface for a human operator remotely controlling UCAVs should strike a proper balance between emotional and moral attachment and detachment. This requires an ethical design of the computer systems used by human operators to make life-and-death decisions without removing the moral-psychological barriers to killing. On the one hand, such systems should communicate the moral reality of the consequences of the decisions of human operators, and on the other hand such systems should reduce the strong emotions human operators feel to reduce the number of war crimes. To develop such systems is a real challenge, but the existence of such systems is necessary to solve the problem of the attribution of responsibility in order to fight a just war.

As we have argued, there can be no value sensitive design for autonomous military robots, since the value of moral reflexivity is a necessary condition for ethical decision-making, which cannot be delegated to nonhumans such as robots (in the near future). We are in favor of the idea of Asaro (2009) to improve further ethical decision-making of the human operators with the help of technology instead of improving robot performance in decision-making.

Where does the above leave the concept of VSD when focused on the value of preventing of innocent casualties in the military domain? Clearly, the noncooperative nature of the domain calls for a wider approach in comparison with VSD applied in a relatively benign, cooperative context. However complex it may be, the search for and a VSD-based analysis and design of NLW technologies and concepts should be complemented by value sensitive scenarios and human behavioral models, in order to arrive at well-balanced and realistic designs and associated applicability assessments.

## Cross-References

► Design for the Value of Responsibility
► Design for Values in the Armed Forces: Nonlethal Weapons and Military Robots

[13]As, for instance, Wright (2006, pp. 190–191) found that during the Troubles, strong correlations existed between events with baton round use and the occurrence of violence and insurgency activity against British Army personnel soon after the such events.

## References


Allgood M (2009) The end of US military detainee operations at Abu Ghraib. Master thesis, University of Florida, Orlando

Amnesty International (2004) United States of America: excessive and lethal force? (Report AMR 51/139/2004). Amnesty International, London

Anderson M, Anderson S (2007) Machine ethics: creating an ethical intelligent agent. AI Mag 28 (4):15–26

Arkin R (2009) Governing lethal behaviour in autonomous robots. Chapman & Hall/CRC, London

Arnesen O, Rahimi R (2007) Military non-lethal solutions for medium to long ranges. Paper presented at the 4th European symposium on non-lethal weapons, Ettlingen, 21–23 May

Asaro P (2007) Robots and responsibility from a legal perspective. In: Proceedings of the 8th IEEE 2007 international conference on robotics and automation. Workshop on RoboEthics, Rome, 14 April 2007

Asaro P (2009) Modeling the moral user. IEEE Technol Soc 28(1):20–24

Bandura A (1999) Moral disengagement in the perpetration of inhumanities. Pers Soc Psychol Rev 3(3):193–209

Burrows C (2002) Operationalizing non-lethality: a Northern Ireland perspective. In: Lewer N (ed) The future of non-lethal weapons – technologies, operations, ethics and law. Frank Cass, London, pp 99–111

Centre PF (1996) In the line of fire – Derry July 1996. Pat Finucane Centre, Derry Londonderry

Coker C (2001) Humane warfare. Routledge, London

Cummings M (2004) Creating moral buffers in weapon control interface design. IEEE Technol Soc Mag 23:28–33

Cummings M (2006) Automation and accountability in decision support system interface design. J Technol Stud 32(1):23–31

Detert J, Treviño L, Sweitzer V (2008) Moral disengagement in ethical decision making: a study of antecedents and outcomes. J Appl Psychol 93(2):374–391

Donnelly S (2005) Long-distance warriors, Time Magazine, 4 Dec

Fieser J, Dowden B (2007) Just war theory. The internet encyclopedia of philosophy. http://www.iep.utm.edu/j/justwar.htm

Fitzsimonds J, Mahnken T (2007) Military officer attitudes towards UAV adoption: exploring institutional impediments to innovation. Jt Force Q 46:96–103

Freedman L (1998) The revolution in military affairs. Oxford University Press, London

Gompert D, Johnson S, Libicki M et al (2009) Underkill – scalable capabilities for military operations amid populations. RAND Cooperation, Arlington

Gulam H, Lee S (2006) Uninhabited combat aerial vehicles and the law of armed conflicts. Aust Army J 3(2):123–136

Gwynn C (1934) Imperial policing. MacMillan, London

Hellström T (2013) On the moral responsibility of military robots. Ethics Inf Technol 15 (2):99–107

Human Rights Watch (2012) Egypt: Protesters' blood on the military leadership's hands. http://www.hrw.org/news/2011/11/22/egypt-protesters-blood-military-leadership-s-hands. Accessed on 12 Feb 2012

Hussey J, Berry R (2008) When the earth shakes! Detainee disturbances in an internment facility. Mil Police 19(1):9–12

Iraq Coalition Casaulty Count (2008) Deaths caused bu IEDs and U.S. deaths by month. http://icasualties.org/oif/IED.aspx. Accessed on 12 Feb 2012

Isenberg D (2007) Robots replace trigger fingers in Iraq, Asia Times Online. http://www.atimes.com/atimes/Middle_East/IH29Ak01.html

Johnson A, Axinn S (2013) The morality of autonomous robots. J Mil Ethics 12(2):129–141

Kaag J, Kaufman W (2009) Military frameworks: technological know-how and the legitimization of warfare. Camb Rev Int Aff 22(4):585–606

Kaplan R (2006) Hunting the taliban in Las Vegas. Atlantic Monthly 4 Aug
Kenyon H (2006) Israel deploys robot guardians. Signal 60(7):41–44
Krishnan A (2009) Killer robots. Legality and ethicality of autonomous weapons. Ashgate, Farnham
Latham A (1999) Re-imagining warfare: the "revolution in military affairs". In: Snyder C (ed) Contemporary security and strategy. MacMillan, London
Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6:175–183
Mayer C (2007) Non-lethal weapons and non-combatant immunity: is it permissible to target non-combatants? J Mil Ethics 6(3):221–231
McNab R, Scott R (2009) Non-lethal weapons and the long tail of warfare. Small Wars Insurg 20(1):141–159
Millar R, Rutherford W, Johnston S et al (1975) Injuries caused by rubber bullets: a report on 90 patients. Br J Surg 62:480–486
Morris J (1992) Non-lethality: a global strategy (white paper). US Global Strategy Council, Washington, DC
Murphy M, Merrit J, Mason P et al (2003) Bioeffects research in support of the active denial system (ADS): a novel directed energy non-lethal weapon. Paper presented at the European Working Group NLW 2nd symposium on non-lethal weapons, Ettlingen, 13–14 May
National Research Council (2005) Autonomous vehicles in support of naval operations. The National Academies Press, Washington, DC
Northern Ireland Office (2006) Patten report recommendations 69 and 70 relating to public order equipment. A research programme into alternative policing approaches towards the management of conflict. Fifth report prepared by the UK Steering Group led by the Northern Ireland Office, in consultation with the Association of Chief Police Officers. Northern Ireland Office, Belfast
Orbons S (2010) Do non-lethal weapons license to silence? J Mil Ethics 9(1):78–99
Orbons S (2012) Non-lethality in reality. A defence technology assessment of its political and military potential. PhD thesis. University of Amsterdam. http://dare.uva.nl/record/436342
Orbons S, Royakkers L (2014) Non-lethal weapons: striking experiences in a non-cooperative environment. Int J Technoethics 5(1):15–27
Riley J, Strater L, Chappell S et al (2010) Situational awareness in human-robot interaction: challenges and user interface requirements. In: Jentsch F, Barnes M (eds) Human-robot interaction in future military operations. Ashgate, Burlington
Royakkers L, Van Est Q (2010) The cubicle warrior: the marionette of digitalized warfare. Ethics Inf Technol 12:289–296
Schulzke M (2013) Autonomous weapons and distributed responsibility. Philos Technol 26(2):203–219
Sharkey N (2008) Cassandra or false prophet of doom: AI robots and war. IEEE Intell Syst 23 (4):14–17
Sharkey N (2010) Saying "no!" to lethal autonomous targeting. J Mil Ethics 9(4):369–383
Singer P (2009) Wired for war: the robotics revolution and conflict in the twenty-first century. Penguin, New York
Smith R (2006) The utility of force – the art of war in the modern world. Penguin, London
Sparrow R (2007) Killer robots. J Appl Philos 24(1):62–77
Sparrow R (2009) Building a better warbot: ethical issues in the design of unmanned systems for military applications. Sci Eng Ethics 15(2):169–187
Sparrow R (2011) Robotic weapons and the future of war. In: Tripodi P, Wolfendale J (eds) New wars and new soldiers: military ethics in the contemporary world. Ashgate, Farnham, pp 117–133
Strawser B (2010) Moral predators: the duty to employ uninhabited aerial vehicles. J Mil Ethics 9(4):342–368

Sullins J (2010) RoboWarfare: can robots be more ethical than humans on the battlefield? Ethics Inf Technol 12(3):263–275
Tanielian T, Jaycox L (eds) (2008) Invisible wounds of war: psychological and cognitive injuries, their consequences, and services to assist recovery. RAND Corporation, Santa Monica
Toffler A, Toffler H (1994) War and anti-war – survival at the dawn of the 21st century. Warner, London
UK Steering Group (2004) Patten report recommendations 69 and 70 relating to public order equipment: a research report into alternative policing approaches towards the management of conflict. Fourth report, Northern Ireland Office, Belfast
UK Steering Group (2006) Patten report recommendations 69 and 70 relating to public order equipment: a research programme into alternative policing approaches towards the management of conflict. Fifth report, Northern Ireland Office, Belfast
US Department of Defense (2009) Unmanned systems roadmap: 2009–2034 (OMB No. 0704–0188). Department of Defence, Washington, DC
United States Air Force (2009) Unmanned aircraft systems flight plan 2009–2047. Headquarters, United States Air Force, Washington, DC
US Army Surgeon General's Office (2006) Mental health advisory team (MHAT) IV: Operation Iraqi Freedom 05–07, 17 Nov 2006. www.globalpolicy.org/security/issues/iraq/attack/consequences/2006/1117mhatreport.pdf
Van de Poel I, Royakkers L (2011) Ethics engineering and technology. Blackwell, Oxford
Veruggio G, Operto F (2008) Roboethics: social and ethical implications of robotics. In: Siciliano B, Khatib O (eds) Springer handbook of robotics. Springer, Berlin, pp 1499–1524
Vilke G, Chan T (2007) Less lethal technology: medical issues. Polic Int J Police Strateg Manage 30(3):341–357
Von Clausewitz C ([1831] 1984) On war. In: Howard M, Paret P (eds) On war. Princeton University Press, Princeton
Weir S (1983) No weapon which deters rioters is free from risk. New Soc :83–86
Wickens C, Levinthal B, Rice S (2010) Imperfect reliability in unmanned air vehicle supervision and control. In: Barnes M, Jentsch F (eds) Human-robot interaction in future military operations. Ashgate, Burlington
Wright S (2006) A system approach to analysing sub-state conflicts. Kybernetes 35(1/2):182–194

# Design for Values in Economics

Aad Correljé, John Groenewegen, Rolf Künneke, and Daniel Scholten

## Contents



**Abstract**

We distinguish the subjectivist value theory, providing the basis for neoclassical economics (NCE) and new institutional economics (NIE), and the social theory of value, underscoring original institutional economics (OIE). In NCE design involves a comparison of the structural characteristics and conduct in a real world market with the theoretically ideal competitive market. If any problems are identified, corrective action should (re)establish competition. NIE would additionally examine the characteristics of the transactions and potential external effects and evaluate the adequacy of the prevailing market institutions. The focus is on designing the right institutions for markets to reveal the

A. Correljé (✉) • J. Groenewegen • R. Künneke • D. Scholten
Delft University of Technology, Delft, The Netherlands
e-mail: a.f.correlje@tudelft.nl; j.p.m.groenewegen@tudelft.nl; r.w.kunneke@tudelft.nl; d.j.scholten@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_24

subjective values of actors. OIE looks at design in a dynamic, holistic, and systemic way and considers "the market" as one of the possible tools out of many to realize social (moral) values. Individual and common values emerge and are constituted in interaction, being judged and deliberated in their specific context of time and place regarding their consequences for society.

**Keywords**
Subjectivist value theory • Social theory of value • Neoclassical economics • New institutional economics • Original institutional economics

## Introduction

In this chapter we will address the notion of design for values in economics. We will discuss how the notion of value is conceptualized in two different paradigms in economics, with different implications regarding the activity of design. Value has been central in economics from the early days onward. In the early history of economic thought, value was supposed to be inherent to the object. Either God had given the object value or man had given it value through its labor, the labor theory of value.[1] The value of goods or services and the exchange rate between them were supposed to be *objectively* determined. With the Marginalist Revolution in the 1870s, the *subjectivist* theory of value was introduced into economics, according to which individual human preferences give value to objects. Individual actors know how they value the goods they want to purchase or produce and express their preferences via their offerings in the market. Value became equated with price. A contrasting approach toward values is advanced in the so-called *social* theory of value that explains how values are constituted in society and why the individual preferences of the "homo economicus" should be replaced by the socially constructed values of the "homo socialis." In the social theory of value, a clear distinction is made between the underlying fundamental convictions actors hold (sometimes addressed as Values with a capital "V") and the economic exchange value of goods and services. Underlying Values are

> (. . .. . .) enduring convictions about what is good, [V]alues can pertain to a good life, to a good or just society, but also to what constitutes a good work of art. (Vermaas et al. 2011, p. 39; see also Dolfsma 2004, p. 48)

This chapter is organized as follows. In the second section, we discuss the subjectivist theory of value. This theory is endorsed in neoclassical economics (NCE) and new institutional economics (NIE), and we explain how these two schools in economics put the individual preferences of the "homo economicus" on center stage. We explain how the related ethics of utilitarianism together with

[1]Adam Smith and Karl Marx are known because of their labor theories of value. For the differences between them and the problems of the theory in general, see Heilbroner (1988).

the so-called deontological ethical rules determine the "free market" as the arena where the individual subjective values can and should be revealed and where they (should) coordinate transactions. In this section we also address the implications of the subjectivist value theory for the issue of design. The section closes with a discussion on how economists design, to realize general values like welfare and specific values like, for instance, privacy and safety. We will use the process of the EU energy market development to illustrate how these subjectivist notions of design are applied in real world economic policy making. The third section addresses the social theory of value, associated with the school of original institutional economics (OIE). We discuss how values are constituted and how values are emerging and designed in an evolutionary, social process. It will be shown that not only deontological ethical rules are of importance but also the more contextualized virtue ethics. This turns out to have large implications for the design issue and likewise for the question how to design for values, as will be shown in our examination of the development of a common energy market in Europe.

## The Subjectivist Theory of Value[2]

### Neoclassical Economics

In the models of standard neoclassical economics (NCE), individual consumers are assumed to decide on the basis of their utility function, the subjective value. The utility of a good or service is the capability to satisfy individual wants, and the value of an object is what the individual ascribes to it because of his/her preferences. Given their preferences, the individual consumers are assumed to maximize their utility through their demand in the market.[3] Individual producers determine what they offer in the market, taking into consideration the profits they collect at a given market price. The aggregate demand of all individual consumers, in confrontation with the aggregate supply, results in a certain amount of goods traded at the equilibrium price, i.e., the price that consumers are willing to pay and that producers accept.

So the individual with exogenously given preferences is the starting point for the economist. This approach, based on individual utility functions, allows for individuals having moral preferences, like acting in the interests of others (Becker 1996). Yet that is outside the domain of economic inquiry. In that sense, according to NCE, economics is a "value-free" science: it does not study and evaluate the subjective values as such, but takes the "revealed" preferences as a given.

[2]This section draws extensively on Groenewegen (2011, 2013).

[3]The consumer continues to demand units of a good or service until all marginal utilities are equal and she is not able to improve his/her total utility anymore.

These preferences are to be revealed in the offerings of the actors as buyers and sellers in the market. The actors in NCE are modeled with specific rules of behavior; they maximize utility and profit and minimize costs. They also have characteristics of full rationality, which makes an ex ante calculation of optimal combinations possible. So the *homo economicus* is a fully informed actor, who is positioned in a well-defined environment of a specific market structure, like a perfectly competitive market, or a monopoly.

This environment is analytically considered a static given, exogenous to the model. In applying the Methodology of the Scientific Research Program on NCE, Latsis (1976) identified the hard core, protective belt, and heuristics of neoclassical economics. He concluded that the core models are all of a "single-exit structure." Given the characteristics of the actors and their situation, logically they have no other option than to calculate and to "choose" the one optimal solution, which is the one theory predicts. This does not only hold for the model of pure and perfect competition but also for the monopolistic and oligopolistic models with well-defined price or quantity reactions.

### Neoclassical Economics and Value

Economists present their discipline as a value-free science in the sense that they do not normatively appraise the subjective values of the actors and in the sense that in their scientific investigation they have objectively access to the facts. The neoclassical theory adheres to the positive-normative dichotomy that separates fact ("what is") from value ("what ought to be"). The wants and subjective valuation of actors are exogenously given objective facts for the scientific researcher. A normative analysis of those facts cannot and should not be part of the economists' scientific inquiry. The positive and the normative should be carefully separated and then, it is claimed, economics is a value-free science. A related tenet to this separation is the claim that the facts are objectively accessible through our senses. The facts economics is studying are "brute facts," i.e., they are in no way constructed by the theoretical concepts applied (see section "Original Institutional Economics (OIE)" below for explanation).

It is on this basis that NCE claims to embody the principle of so-called "ethical relativism" (Tool 1986); all values, criteria, and preferences are relative to individuals. It is considered inappropriate to judge the values and criteria on the basis of which the rational individuals choose. Such wants are given and "utility" is taken to be a proxy for Values. "Thus utility is the meaning of value in orthodox neoclassicism and price is its measure" (Tool 1986, p. 9).

In NCE the original Benthamite utilitarian principle of comparing individual utility is considered impossible. It is replaced by the Paretian principle, stating that one can only identify situations in which society as a whole is better off. Because comparing the utility of individuals is impossible, a redistribution between individuals is rejected. The Pareto optimum states that the situation is improved when at least one person is made better off without making anyone else worse off "(. . ...) thus removing the moral basis of utilitarianism from welfare economics (. . ..). By so absorbing morality into subjective and incomparable individual preferences,

neoclassical economics has effectively removed ethical evaluation from welfare analysis" (van Staveren 2007, p. 22).

## New Institutional Economics

Since the mid-1970s, the school of new institutional economics (NIE) developed strongly with Nobel laureates like Ronald Coase (1991), Douglass North (1993), Elinor Ostrom, and Oliver Williamson (2009). The earlier Nobel laureates Friedrich von Hayek, Kenneth Arrow, and Herbert Simon are often considered to be institutionally oriented economists (Williamson 1975).

The NIE addresses questions that were no part of NCE, such as why do institutions like property rights and firms exist? What is the role of values and norms in society? What is the impact of differences in the institutional environment of economies, markets, or sectors on the allocation of goods and services? In short, why do institutions exist and why do they matter? In addressing such questions, NIE introduced two additional attributes to the economic actor: *bounded rationality* and *opportunistic behavior*. The first is about the limited capacity of actors to capture all relevant information and to calculate their individual optimal outcome or to make a complete contract in which all eventualities are taken care for. The second is about the possibility that actors abuse asymmetry of information, by providing misleading information to others or even by cheating them.

Hence, NIE positions the actors in complex and uncertain environments implying that they are not able, as in NCE, to eliminate all uncertainties through complete contracting. So, as is argued, to govern their transactions in an efficient way, the actors create institutional arrangements like vertically integrated firms, a variety of (long-term) contracts, forms of cooperation, and branch associations. Maintaining the value-free philosophical and methodological characteristics of NCE,[4] NIE explains that institutional arrangements exist because they are efficient in minimizing transaction costs. Therewith, the actors are able to reduce their overall cost of supplying and purchasing the particular good or service in the market. Hence, the transactions take place at lower prices enhancing overall societal welfare.

Williamson (1998) presents his view on values and how different schools of economics deal with value in Fig. 1. At level 1 NIE explicitly recognizes the existence of values reflected in informal institutions as exogenous variables of importance. These informal institutions include attitudes, norms, customs, and religion.[5] Informal institutions are not explicitly formulated and written down but are internalized in the hearts and souls of the members of a community. Informal

[4]A distinction is made between the so-called Williamsonian and the Northian branch of NIE (Groenewegen 2011). In our interpretation we conclude that the former stays in the philosophical and methodological tradition of NCE, whereas the latter departs from it and adopted many characteristics of the original economic institutionalists (see below).

[5]Williamson (1998) locates those informal institutions at level 1. He does not explicitly distinguish values. We consider values to be at level 1.

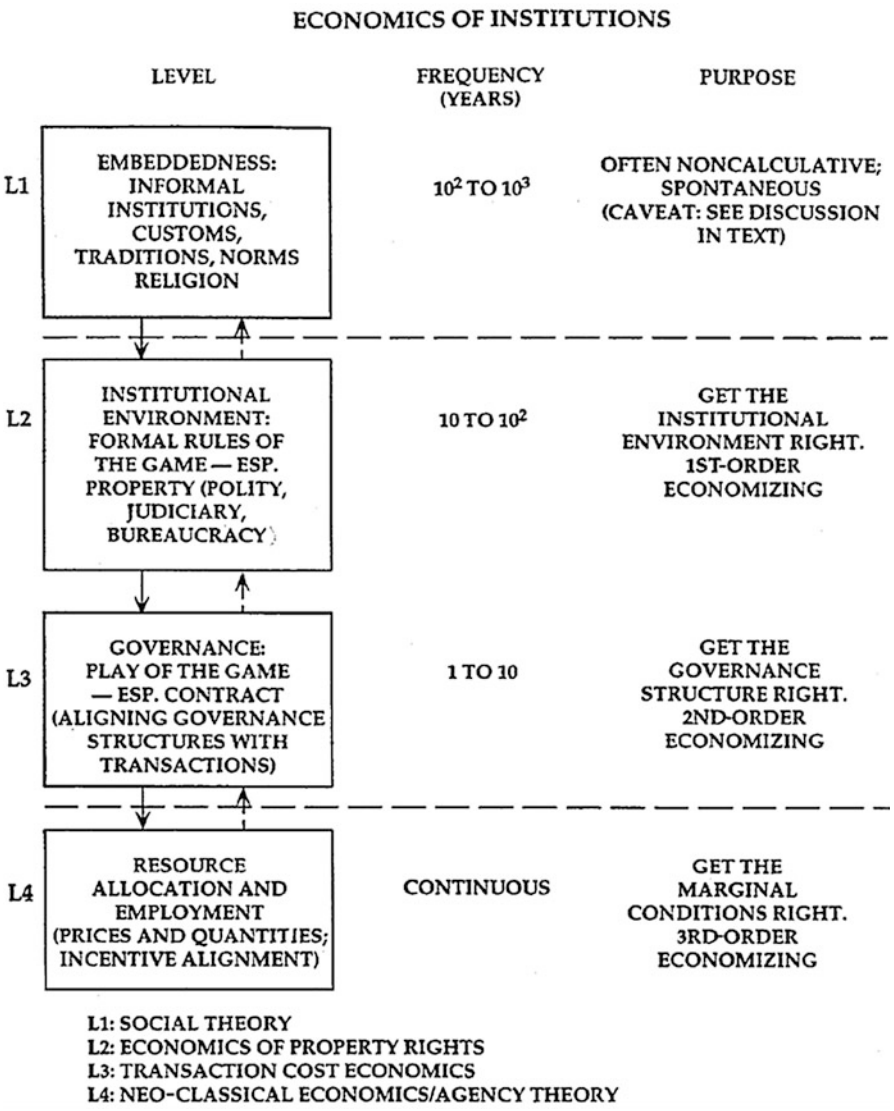



**Fig. 1** The Economics of Institutions, from Williamson (1998, p. 26)

institutions change slowly and are not subject to economic calculative behavior. Individual economic actors, and even groups of actors in collective action, are rarely in the position to purposefully (re)design informal institutions. Informal institutions emerge "spontaneously" out of the interactions of millions of actors (see below the discussion of Aoki 2001, 2007). Informal institutions are in the domain of "social theory" and considered a given for the economist. Yet whereas they are no object of economic analysis, it is important for an NIE economist to be

well informed about the informal institutions in a country, market, industry, or firm, because they have a strong influence on the formal institutions at level 2.[6]

The formal institutions, the "rules of the game," like laws and regulations belong to the domain of NIE and are subject to so-called "first order economizing." The central theory at level 2 is the theory of property rights. Different configurations of property rights (private, public, collective, and common) influence the behavior of actors differently and produce different outcomes. Societies that aim for efficient allocation of their scarce resources better "get their formal institutions right," including the agencies that monitor behavior and enforce rights. Clear, enforceable property rights, an independent judiciary, and an objective bureaucracy should be designed to provide individual actors with the right incentives to maximize their profit and utility or to minimize costs.

Level 3 refers to the next step on the path of "economizing," the governance. The purpose of transaction cost economics (TCE) is to understand why different modes of governance exist to coordinate economic transactions. TCE explains the existence of different types of contracts of hierarchical organizations like firms, multidivisional firms, and multinational corporation, of regulatory agencies, and of state-owned enterprises.

Williamson became a Nobel laureate (together with Elinor Ostrom) for his work on governance. He showed how specific governance structures could be matched, "aligned," with specific types of transactions so that transaction costs are minimized. Transactions in a market can be distinguished by particular characteristics. A chief aspect here is the degree to which the assets involved are specific to a transaction, locking in the economic actor, or whether they can be (re)employed easily in other uses (see below). Other important issues are the frequency of the transactions taking place, uncertainty in the market, and the nature of surrounding informal and formal institutions and of the actors involved. These are the independent variables indicating a greater or lesser need for the actors to make safeguards against their potential opportunistic behavior. The dependent variable is the governance structure; the larger the possibility of opportunism, the more complex the contract will be, the higher the transaction costs involved, and the more efficient a more hierarchical arrangement may be. Eventually, above a certain threshold of transaction costs, private actors may withdraw from the market, and it may even be necessary to invoke public oversight or state-owned enterprise to carry out certain economic activities.

In new institutional economics, the line of reasoning of neoclassical economics is maintained: the actors maximize profits and utility and minimize costs. They do so in the prevailing environment of formal and informal institutions. An important underlying assumption is about the selecting role of competition. The competitive

[6]The arrows in Fig. 1 indicate a causal relationship between institutional environment and governance structure; the dotted feedback arrows indicate that "Although, in the fullness time the system is fully interconnected, for my purposes here, these feed backs are largely neglected" (Williamson 1998, p. 26).

market is supposed to force actors to select the most efficient governance structure; otherwise they will not survive the rivalry with their competitors. The economic allocation and prices and quantities transacted by the actors at level 4 then are an issue for NCE and principal-agent theory[7].

## From Comparative Static to Dynamic

The TCE approach à la Williamson is in essence a comparative static approach, in which alternative governance structures are evaluated against each other on the basis of their efficiency (Groenewegen and Vromen 1996, Groenewegen and de Jong 2008). Within NIE, Masahito Aoki has developed a dynamic approach of institutional change that fits well with the NIE characteristics of efficiency and equilibrium. In his "comparative institutional analysis," the emergence and evolution of institutions are not explained as the result of a purposeful collective decision but as the unintended result of a sequence of micro decisions. To understand the spontaneous emergence of institutions, the so-called coordinated game is relevant. When the domain of the game is specified in economic, political, judicial, or social terms, while the choices the agents can make are specified as well, then it can be shown that boundedly rational actors will create stable institutions over time. According to Aoki (2007, p. 7), "An institution is a self-sustaining, salient pattern of social interactions, as represented by meaningful rules that every agent knows and are incorporated as agents' shared beliefs about how the game is played and to be played." Aoki conceptualizes institutions as equilibria, which emerge out of the unintended actions of the actors at a decentralized level in the system. Accordingly, he calls his approach the "institutions-as-an-equilibrium approach." How do these social rules come about?

"Institutions are the result of human action, but not of human design" is a well-known saying that captures the essence of the evolutionary approach. This refers to actors behaving in a specific way because it is in their own interest to do so and as an unintended outcome an institution like a norm or convention emerges. That behavior at individual level of the actors can be intentional (they aim, for instance, at minimizing costs) or routinized (not being aware, actors follow a specific rule). The point in the evolutionary approach is that institutions can come about without individual or collective action intended to create the institution or to change the existing one. The outcome of all the individuals' behavior can be the emergence of an institution that, once in existence, is durable and structures individual preferences, behavior, and social interaction (see also Greif 2006). The explanation of the emergence of such durable institutions is efficiency based. All actors consider behavior in line with the emerging norm to be in their own interest and would

[7]The positioning of agency theory (AT) only at level 4 is confusing: Elsewhere (Groenewegen et al. 2010) is explained that the positive AT can be best located at level 3 and the normative one at level 4.

like to see others to behave likewise. The actors discover that it is costly to ignore the emerging institution and that it is beneficial to follow suit, therewith reducing uncertainty and information costs. This insight is growing over time when the institution develops and actors increasingly share the same knowledge, establishing the institution and creating an equilibrium.

According to the "institutions-as-an-equilibrium" approach, the regularity has become an institution, a norm, or a convention, when a large majority of the actors in the community have internalized the regularity. In the literature this approach is also called "spontaneous" (see level 1 in Fig. 1), because it is a matter of self-enforcement. No external authority forces the actors to behave according to a specific institution. The process emerges spontaneously purely based on the self-interests of the individuals.

## The Design Issue in the Subjectivist Theory of Value

According to the subjectivist theory of value in neoclassical and new institutional economics (both in the Williamsonian and the Aokian version), individual actors should be able to reveal their subjective values or preferences in an efficient market. Consequently, the design issue is about the shape of markets or more broadly the design of market economies, which include the political and social institutions that support the market. When the market is designed well, automatically the best possible outcome will result.

### The Design Issue in Neoclassical Economics

The design issue in NCE focuses on the value of efficiency in market structures, as a means to let individual consumers fulfill their subjective values. NCE adherents are convinced of one thing: competition will bring the best outcome possible for society, whatever that may be. A well-functioning market will reveal the outcomes over time, and prices will reflect the aggregated preferences of the consumers and the optimal combinations of the production factors.

When the specific conditions concerning the market structure and formal competition law are fulfilled, competition will put the individual suppliers in the market under pressure, resulting in a search for the most efficient combinations of scarce resources, to offer the individual consumers products and services at the lowest price possible, fulfilling their subjective values. Moreover, the market will push participants to innovate, in products, production processes, and governance structures, to stay ahead of competitors. So subjective values are central and the market is the most effective and efficient means to realize society's welfare. Note that in NCE the introduction of the market and its competitive nature does not tell us anything about the specific outcomes the market will bring us. Which services will be offered, at what prices, and which values of whom will be fulfilled are unknown, ex ante.

In NCE this conviction has led to a description of the ideal type of a competitive market with many independently operating suppliers and many consumers as the

reference model and to a corresponding competition policy. All market actors should have equal access rights to input and output markets and to the relevant objective information. All private institutional arrangements, like vertically integrated firms, strategic alliances or acquisitions, and public interventions other than based on competition law, are considered "anticompetitive." This is because they reduce the number of market participants, their independent behavior, and the private choices of the trading actors. Firms are production functions and the market is a "signaling device," in which prices objectively signal consumers and producers what the scarcities are and when these have changed due to exogenous shocks. Also the consumers in the mainstream market economy are assumed to serve the functioning of the market economy, and the right consumer behavior is to switch to another product, or supplier, when price and quality differences indicate so.

The universal (moral) rules of the (competitive) game described above and the corresponding rights and norms of consumers and producers in the market are part of the so-called deontological ethical rules that embed the neoclassical market and that should be enforced by legal measures to make the market function properly (Van Staveren 2007).

An illustrative application of this perspective on a process of market design is provided by the development of the Internal Energy Market in Europe, post-1988 (CEC 1988). This policy was enacted by the European Commission with the predominant objective to establish a single market for energy within the area covered by its member states. To this end the Commission has launched a variety of initiatives to reduce the barriers to trade within and between the member states for coal and oil products as well as for grid-bound electricity and natural gas. Main objectives were the dismantlement of the large variety of national regulatory structures and public and private monopolies through which the trade in fuels and power was coordinated traditionally. Open, EU-wide markets should be created to allow consumers to select the best offers from competing suppliers. The main tools available to the Commission were, on the one hand, the traditional instruments of competition policy and, on the other, instruments of sector specific regulation, formulated in Electricity and Gas Directives and to be implemented by the member states in their legal frameworks (CEC 1996).

In essence, the idea was that the energy industries would require price and access regulation over a transitional period, limited in time, after which a competitive market would have been established, taking care of an efficient coordination of the transactions between buyers and sellers. Consumers would follow their preferences and select the most adequate suppliers, taking into consideration prices and quality. Governments would withdraw from intervening in the markets, the overall costs of energy supply would fall, quality and services would improve, and excessive monopoly rents would disappear. Energy would become a normal commodity, it was argued.

## The Design Issue in New Institutional Economics

Expanding the world of NCE, Williamson (1975) showed how all kinds of private governance structures were not only meant to create market power but also aimed to

reduce the transaction costs. Hence, they should not always be forbidden by competition law; the subtitle of his book was "implications for antitrust policy." Private ordering of markets serves efficiency according to NIE, not only to be calculated as a minimization of production costs as in neoclassical economics but also as the minimization of transaction costs.

NIE introduced another economic vision on private and public institutional arrangements, and Williamson (1979) discussed a range of efficient governance structures. Each of these structures can be efficient to coordinate specific types of transactions, depending on the degree of asset specificity of the good or service transacted. As stated above, when the investments are very specific then the asset is worthless when the transaction is ended. The degree of asset specificity has implications for the possibility of opportunistic behavior and therefore for the need of safeguards.

When transactions have a low asset specificity, then the "ideal" traditional (spot) market contract is most efficient. Such transactions involve either (low) capital investments that can be made productive in other applications without costly adjustments, or they provide more or less standardized goods and services that can be traded easily. Hence, the investor or producers are not locked into a transaction with a specific purpose, at a specific location, or with a specific partner. An illustrative example is the trade in oil products, with a large market for more or less standardized fuels, like gasoline, diesel, and kerosene, being produced in refineries that accept a wide variety of crude oil inputs, from many producing countries all over the world, being delivered by ships that can go anywhere. Indeed, the danger of potential opportunism is largely absent, because of the high substitutability of the good and the high level of competition in the market, both among the suppliers and the customers.

When investments (assets) are tied up to particular uses, locations, and customers, their asset specificity is said to increase. For example, gas and electricity supply systems involve gas and power production from gas fields and power plants, cables and pipelines transporting the gas and power from A to B and C, and the gas and power distribution networks, connecting the customers, with their specific customized gas and power appliances. But the required hardware assets of the producers, the transporters, and the users are highly "specific" to their function in the system, their location, and often their use in combination with other assets. A windmill without access to a connecting cable is worthless and vice versa. The profitability of huge investments in underground gas storage facilities is jeopardized when the contractors of the capacity are free to shop around among such facilities, driving down the storage tariff to rock bottom levels. So nobody will invest, facing this risk.

Functioning systems require all components to be in place and to be used, at a sufficient level of supply and demand to justify the cost of the installed assets. Certainty of supply and of delivery at an acceptable price, covering the cost and an adequate remuneration, is then to be assured through long-term contracting or so-called "hybrid" governance structures that limit the autonomy of the actors in behaving opportunistically. So when high risk is at stake because of asset

specificity, the efficient governance structure shifts from market to the hierarchy of the (contractually) vertically integrated firm.

Williamson moved from there into the public sphere of regulation and state-owned enterprises ("public bureau"), arguing that at even higher levels of asset specificity and uncertainty, public governance of regulation and state-owned enterprises are required to facilitate such transactions, a prime example being flood defenses in which no one would invest in as a private undertaking. But also the transport infrastructure in power and gas systems is often publicly owned, for the reason of "security of supply."

Hence, the design issue in NIE is about "getting the institutions right" at level 2 and "getting the governance structures right" at level 3 of Fig. 1. Although Williamson directed most of his attention to level 3, many other NIE economists also apply transaction cost insights to design issues at level 2. Spiller (2013), for instance, has demonstrated how the cost of regulation in "public-private contracting," as contrasted with Williamson's private-private contracting, brings in situations of governmental and third party opportunism and shows how institutions can be analyzed and designed to minimize such public and politically inspired opportunistic behavior.

The notions of uncertainty, asset specificity, and opportunistic behavior may explain the evolution of the design process of EU energy market policy post-1990. In contrast to the initial expectations, the process of market restructuring turned out to be highly complex and politically sensitive. Most member states were hesitant in dismantling their prevailing national industry structures. They saw the creation of a competitive industry primarily from the perspective of preparing their own public and private firms, to withstand competition at home while expanding into neighboring markets (Thomas 2003). Moreover, the interests of national energy industries, like the production of coal and natural gas or the nuclear technology complex, were defended both economically and strategically under the banner of "national interest." Interestingly, under the banner of "consumer protection," end-use prices became regulated explicitly in quite a few countries (Haaland Matláry 1997).

In response, in 2003 and 2009, the Commission strengthened its sector regulation efforts by issuing new Directives with increasingly far-reaching requirements, taking into consideration the character of the transactions at stake and the environment of the system in which that took place. These involved, firstly, the vertical unbundling of the national or regional supply monopolies in the electricity and gas industry. Gas and electricity transport networks were separated from the potentially competitive gas and power production companies and from commercial trade and retail activities. The networks, as "natural monopolies," were to be regulated on a cost-plus basis as regards their access conditions, tariffs, and investments, to provide "third party access" to all of the trading parties. It, secondly, involved the horizontal unbundling of formerly monopolist firms to create competition in trading, i.e., the national monopolies had to be divided up into a number of firms that were expected to compete with each other. Moreover, the EU Directorate General for Competition began to actually intervene in the market, by prosecuting large, dominant energy firms on the ground of abuse of market power and

hindering competition. Thirdly, the European Commission sought to establish trading platforms in wholesale markets, arrangements for the cross border transport of power and gas, and an EU-wide Agency for the Cooperation of National Energy Regulators (ACER), aiming at the further interconnection of the national markets (CEC 2003, 2009a, b).

Essentially, all such institutions were part of the design of a "well-functioning European energy market version 2.0," taking into account – and correcting – the variations in asset specificity and uncertainty in production, transport, and (retail) trading, as outlined in the NIE theory (see Joskow 2008; Spanjer 2009; Correljé et al. 2012, 2014).

Step by step, the new contours began to appear of a different European energy industry, consisting of a relatively small number of large internationally operating multi-utilities, producing and trading both power and gas, surrounded by a large amount of asset-light firms that traded on the national exchanges. These suppliers and traders were facilitated by national or regional transport and distribution systems that were regulated by newly established national regulatory authorities (NRAs), in respect of their operation, their income, their tariffs, and their investments. Hence, until halfway the first decade of twenty-first century, the priority of the restructuring process had been with NCE- and NIE-inspired market design, to enhance societal welfare and economic efficiency, however, against vested interests in industries, national resource sectors, sustainable energy beliefs, and (unionized) labor (Thomas 2003; see Correljé and De Vries 2008).

## Design for (Moral) Values

In applying the subjectivist theory of value to the design of institutions, implicitly two central assumptions are made, namely, that individual subjectivist value judgments count first and foremost and, secondly, that the market is the institution that offers the most efficient way for individuals to reveal and realize their values (Van Staveren 2007, p. 1). Consequently the design of an effective and efficient market comes first, with deontological ethical rules that allow and facilitate (many) actors to enter the market on an equal basis, that secure that these actors have access to the relevant information, and that take care of fair competition. These rules are for NCE exogenous to the economic analysis, as given constraints that delineate the arena of the so-called free market where actors can make choices according their subjective values. For NIE these rules are part of the economic analysis and, as such, they are subject to "design."

Next to constraints concerning the proper functioning of the market, societies can design additional constraints concerning requirements regarding production, distribution, and consumption, based on moral values. Examples are found in biological engineering (genetic manipulated food), the ban on child labor, environmental care, and the like. According to the subjectivist theory of value, the arena of the free market is then further restricted to the benefit of those collective values a society explicitly wants to realize.

Within the subjectivist perspective, a main argument for publicly intervening in the workings of the "free" market involves the so-called externalities. Externalities are costs and benefits for individual actors that are not reflected in the prices they pay for goods and services. So actors bear costs or enjoy benefits they are not compensated for or for which they do not pay a price. The well-known examples are negative externalities of power plants that pollute the environment. The price of electricity does not reflect the societal cost of air pollution, as long as there is no market that registers the cost of pollution of clean air. The costs of irritating eyes or, even more serious, the costs of health damage to the people living in the neighborhood of the plant are not internalized in the price of electricity.

In subjectivist economics two ways are suggested to "internalize" such externalities. One way is suggested by the work of Pigou and the other is based on the work of Coase. Both the Coasean and Pigovian perspectives internalize the externality via cost-benefit analysis, by objectively calculating the corrections that should inform actors about the "right" prices, resulting in technical and allocative efficient allocation of the scarce resources. In the Coasean world, the individual actors know the price of either the external costs or the benefits they are willing to pay or accept. This forms the basis for their negotiations about changes in the property rights. Hence, design is geared toward creating institutions that allow the individual actors to freely negotiate about private contracts that internalize these external costs and profits (Coase 1960). In the Pigovian world, a higher level of collective decision-making, like a regulatory agency, is established to correct the outcome of the free market by calculating the required corrections of the prices by means of taxes and subsidies. So the externality is objectively internalized in the market price (Pigou 1932).

In both worlds the arena of the market for individual actors to buy and sell is preserved and analytically (and policy wise) extended into the domain of the externality. In the European energy sector, prevailing sustainability issues involved local pollution externalities, like sulfur, lead, NOX, and particles emissions, to be solved by neoclassical instruments as product and process specifications (property rights) and taxation (externality pricing). Specifications, as the outcome of negotiations between "producers" and "victims" under guidance of an authority, determine the acceptability of a specific level of pollution, i.e., the right to pollute or to accept pollution, as the balance between the costs incurred by avoiding this pollution by the producers or the cost of the harm inflicted upon the victims. Taxation may add the assumed societal cost of pollution to the free-market price for a good, therewith adjusting the supply/demand equilibrium and reducing the amount of pollution.

To summarize, in the subjective theory of value the arena of the free market, revealing and implementing subjective values, is constrained by: first, rules that make the free market function properly; secondly, interventions that correct prices in order to internalize externalities; and, thirdly, rules that constrain actors and that oblige and forbid, in order to realize societal (moral) values. Hence, the market is conceptualized as a free arena, yet constrained by a set of rationally set rules that belong to the institutional environment of that market.

## The Social Theory of Value

### Original Institutional Economics (OIE)

In the USA, by the end of the nineteenth century, Thorstein Veblen was a well-known institutional economist, highly critical of neoclassical economics (Veblen 1899, 1904). In his opinion NCE was too formal and abstract and too static and wrongly based on the theoretical assumption of individual actors that are disconnected from their institutional environment. Until around 1945 an influential group of American institutional economists dominated the further development of institutional economics. Wesley Mitchell (1927), John R. Commons (1931, 1934), and Clarence Ayres (1944) joined Veblen in his criticism of NCE and underlined the importance of including institutions in the economic explanation (see Gruchy 1972).

The work of those institutional economists is called original institutional economics (OIE). With respect to values, the OIE developed the so-called social theory of value. Values are not considered to be exogenous to the economy and based on the individual preferences, but they are constituted in a process of interaction between individuals where preexisting values play a structuring role. This is a fundamental contrast between the subjective (NCE and NIE) and the social theory of value (OIE),[8] which incorporates a number of other crucial differences, like the attributes and motivations of actors, the structures that embed actors, and the interaction between actors and structures.

According to OIE, the economy, first of all, is an evolving system, in which actors of a different nature (political, economic, social) with different interests and capabilities and with different amounts of power take decisions. They act, react, follow, initiate, and choose. In doing so, these actors are constrained and enabled by structures such as technology and formal and informal institutions and also by their own "mental maps" (Denzau and North 1994). In the evolving economy actors, structures and values are mutually constituted. The nature of economic reality is one of change, and the core research question economics should pose is first of all about understanding that change.

A second important difference between NCE and NIE on the one hand and OIE on the other is in the nature and role of markets as allocation mechanisms. Above we explained how markets are conceptualized according to the subjectivist theory of value: markets are neutral and prices should reflect subjective values. Intervention from "outside" is allowed to make either markets function properly or when constraints are needed for exogenous (moral) reasons. In line with the social theory of value, OIE approaches markets and nonmarket allocation

[8]Original institutional economics (OIE) was after the emergence of NIE often called old institutional economics. We prefer the terminology of original. The label of neo-institutionalism is also used for the postwar institutionalists like John K. Galbraith, Gunnar Myrdal, and others that followed the approach of Veblen and Commons (see Gruchy 1972). In this contribution we call the pre- and postwar institutionalists both OIE.

mechanisms differently. Firstly, the question about societies' collective values is asked: what *ought* to be and what is the end. Then the actual situation is characterized and analyzed, the *is*. If there is a gap between the *ought* and the *is*, the question how the gap should be repaired arises. When (intrinsic) values (and their related instrumental values and policy objectives) do not match with the actual performance of the economy, how then to intervene? An important starting point of OIE analysis is normative: what are the values of societies to design for (the "ought," the "end") and when these are compared with the "is," what then to do about the gap? That is what OIE economists mean when they claim OIE is problem solving and policy oriented.

In order to comprehend the role of individual and collective actors in the process of change, OIE considers a deep understanding of the drivers and motivations of actors of utmost importance. Institutionalists want to know about the "why," so in case another outcome is desired, they have to know how behavior could be changed, by means of what kind of interventions. Instincts, habits, and customs are seen as important drivers and motivations for human decisions. Habits, for instance, are dispositions of actors that have evolved over long periods of time and form the basis of many of the actor's decisions. It would be a misunderstanding, however, to consider habits as mechanically repeated behavior: "habits of thought" form the foundation of much of our behavior and contain past beliefs and experiences, but at the same time human actors have more or less capacity to deliberate and to choose, depending on their environment. They are also "volitional" (Commons 1934; Bromley 2006). Moreover, actors are able to identify habits, to analyze how these influence behavior, and to evaluate whether the habits contribute in realizing the desired consequences of actions or not. If this is not the case, then actors can make existing habits and their consequences explicit and start a process of deliberation in an attempt to change habits (Bromley 2006; Hodgson 2004).[9]

In the OIE framework, actors are positioned, with evolving "cognitive structures," in an evolving institutional context; actors and structures are mutually constituted. Economic actors are social actors operating in specific institutional environments, while markets are institutionalized structures, in which power is equally important as efficiency to understand their performance. It is, according to

[9]Interesting is the question what room is left for volition, for rational purposeful action. In this respect the distinction between habits and routines becomes important. Dewey (1922, p. 28) explains that habits also can be inquired and tested by man, i.e., man can take distance from the specific habits that cause an action and reflect on the consequences of that action. When such reflections raise doubts about the rightfulness (is the "is" well analyzed?) or desirability of the belief (do the habits contribute to the realization of the "ought"?), then man is in the position to inquire what is wrong about the habits causing the undesirable action and to intervene by altering the institutions (the rules of the game) to change the "habit of thought." In the case of routines, man acts mechanically, without thought about the consequences and without valuation of the consequences of the routinized actions in the light of the societal goals. The real opposition is not between reason and habits, but between reasonable habits and unintelligently routinized habit (Costa and Castro 2011).

OIE, a fundamental misconception to present markets as neutral anonymous selection mechanisms, in which individuals independently decide, as if they were atoms. Markets are political constructs, strongly regulated by informal and formal institutions, the rules of the game. In part, these rules evolve spontaneously, especially the informal ones, but they also result from purposeful design.

Moreover, the political process of institutional design and redesign is heavily influenced by societal interest groups. It is characterized by struggle and conflict, because a change of rules almost always implies an adjustment of the distribution of costs and benefits. Consequently markets are best perceived as evolving systems, in which individual and collective action results in both intended and unintended consequences. As a consequence, markets are never in equilibrium but always in a process of adaptation, transition, and evolution.

## OIE and the Social Theory of Value

The existence and constitution of collective values are explicitly taken on board in the social value theory. On the one hand, values underlie the formal and informal institutions of society, and through that "filter" they determine the (economic) values as terms of exchange (Dolfsma 2004, p. 49). On the other hand, the analysis undertaken by the economist is not value-free; facts are always theory laden and on top of that theories are value laden. In contrast to the subjective theory of value, facts and values are not separate categories. Reality is not considered to be composed of objects or "brute" facts, to which the researcher has direct access and which would allow for having objective knowledge about. On the contrary, in order to understand the complex reality, people in daily life and researchers in scientific inquiry make use of "ordering ideas," like concepts, categories, and frameworks that allow for abstraction and that structure the "brute" facts.[10] The world of facts is complex and continuously data have to be sorted out, applying specific standards of relevance (Bush 2009). In selecting the proper standards, inevitably choices are to be made and then unavoidably values and value judgments are involved.[11] Facts speak as far as they are considered relevant from a specific value point of view, generally embedded in a specific theory.

Consider the example of a price hike for crude oil (Tool 1995). Such an increase in the price can be caused by an increase in demand or a reduction of supply. And then the price change reflects the new scarcity, stimulating actors to consume less oil and to look for substitutes. Yet, alternatively, the price increase can also be seen as a reflection of the use of dominant positions in the market; powerful actors or

[10]"Structuring reality" should not be interpreted as "creating reality."

[11]Bush (2009) makes a distinction between values (standards of judgment), valuation (the application of those standards), and value judgment (the evaluation of values in relation to (other) intrinsic values).

their cartels manipulate the supply in order to increase the price. A third theory might link the price hike to an increase in forward prices, reflecting the higher exploitation cost in the future. Hence, the brute fact of the price hike does not speak for itself. It requires a theory, and consequently facts will be investigated and interpreted from that specific theoretical perspective, involving a particular belief about how such markets function and which forces cause the changes observed.

An analysis from a neoclassical perspective addresses other questions and collects other data than an investigation driven by an institutionalist perspective. In OIE the facts are sorted out based on the theories, concepts, and categories that the analytical framework provides. In addition, the social theory of value never claims that the choice of theory is value-free, as it is always guided by underlying intrinsic (moral) values. The value of egalitarianism will lead the researcher to, for instance, the capability theory, whereas the value of efficiency will lead the theorist to the theory of marginal productivity. Boulding (1969) explains that scientific communities like "schools of economics," similar to all other communities in society, adhere (often implicitly) to specific values and apply specific value judgments, guiding the selection of what they consider "appropriate" theories. As a consequence, facts are theory laden and theory is value laden.

In sum, OIE works with a framework that addresses institutional issues in a dynamic, holistic, and systemic way (Wilber and Harrison 1978). In doing so, actors in theories and models are not one-dimensionally efficiency driven, but their preferences are endogenously constituted in the process of interacting and acting.[12] Correspondingly the environment is not only complex as in NIE, but the structures in the environment are constituted mutually with the individuals and collectivities. In contrast to the methodological individualistic approach of the subjective theory of value, the social theory of value is characterized by so-called methodological interactionism, including both the interaction between actors and structures and the interaction among actors.

To put it differently: all values, both individual and common, are constituted in interaction and values both emerge and are designed. Values can be right or wrong, they are subject to (e)valuation, and they are judged and deliberated in a specific context of time and place. The social theory of value is about the social construction of values and about the social processes of judging values. To judge, values are investigated on their consequences for the well-being of the members of the society: what are the consequences of implementing specific values for realizing other more fundamental values?

[12]This is the core of philosophical pragmatism. In the words of Nooteboom (2013, p. 2), pragmatism "(. . .) holds that cognition, in a wide sense that includes normative judgments and goals, occurs on the basis of mental dispositions and categories that are developed in interaction with the physical and especially the social environment." The crux of the argument is that action, practice, constitutes the actor: "Intelligence is internalised practice." This connects well with the framework of North (2005) about institutional change.

## The Design Issue in the Social Theory of Value

Because values are contextual and dynamic, the social theory of value designs institutions that make a "social construction" of values possible, in such a way that individuals in the process of deliberation a) have access to the necessary information, b) have access to the arenas where the deliberation and decision-making takes place, and c) can participate and also have the capabilities to do so in a responsible way. Indeed, actors should be informed, knowledgeable, and aware of their responsibilities.

In design for values, both markets and nonmarket institutions enable individuals to reveal their endogenous preferences and values and offer ways to decide about collective values. It is not only about "free markets" where individuals express their subjective values but also about rules of the game on how collective values ought to be "revealed and implemented." Moreover, the so-called virtue ethics is part of the social theory of values. Local, contextual virtues of actors should be made explicit and are also subject to judgment; some virtues are more "right" than others. Moreover, this judgment may shift over time.

### The Value of Sustainability in EU Energy Supply

Going back to our example of European energy policy making, it is evident that the NCE/NIE-based perspective of efficiency-driven market creation and facilitation was predominant until halfway through the first decade of the twenty-first century. Creating a well-functioning market appeared highly sensible and "right" in a period of a seemingly abundant energy supply and low oil prices of about 20 US dollar per barrel. In the course of that decade, however, European and national policy makers were confronted with new challenges, as regards the shape of the sustainability issues when the origins and threats of global warming were gaining credibility, as well as concerns for the security of energy supply and, more recently, the societal acceptability of new ways of energy production, like wind and solar power and shale gas. In the meantime, the traditional issue of economic efficiency – operationalized as low-cost energy supply as a precondition for economic growth – did not disappear from the policy agendas. Moreover, it also became apparent that the potential solutions to any one of these challenges often would jeopardize the achievement of the other objectives (Pérez-Arriaga 2013; Glachant et al. 2013).

Apparently, the efficiency driven aims to establish a competitive energy market of the 1990s up until 2005 only covered part of the values considered important by European publics and policy makers. The response to global warming involved a call for the reduction of carbon emissions by moving away from fossil energy use, as a new objective for EU energy policy. The value of decarbonization was to be internalized in the economy, alongside the instrumental values of efficiency and competition. The solution created was the EU ETS, introducing a market for trading a capped amount of carbon emission rights among the main groups of users of fossil energy. For reasons of transaction costs, i.e., the measurement of their scattered emissions, residential consumers were kept out of the scheme. For reasons of

international competitiveness, i.e., political power and influence, large energy intensive industry was exempted. So far, however, the ETS malfunctions as its budget of emission rights was established and allocated rigidly in the expectation of continuous economic growth. The economic downturn post-2008 created an overhang of emission rights, driving the price of the certificates down to a quantité négligeable and reducing the effects of the scheme in stimulating green energy to zero.

To a number of EU member states, the disappointing achievements of the ETS provided the "right" reason to implement other instruments to achieve their sustainability objectives. Hence, a variety of support schemes for (de)central wind and solar power generation was implemented in the EU countries. Often such national schemes were calibrated to serve the interests of the industries, local communities, and actors involved in countries, like Germany, Spain, Denmark, the UK, etc. Yet, in most of the countries that were really effective in expanding the supply of green energy, the schemes are now being revised or even withdrawn. The main reason is the fact that funding is becoming prohibitively expensive for the subsidizing states and/or for the energy consumers that have to foot the bill. Nevertheless, (process) innovation and economies of scale have made solar panels and windmills increasingly affordable. So they will continue to be installed.

The obvious success of solar and wind energy is also threatening other important social values. Reliability of energy supply is jeopardized by the impact on power grid stability and balancing. The locations of power generation are generally not very close to the centers of consumption. So additional high-tension power lines have to be constructed throughout countries or to connect offshore wind farms. Lack of societal acceptability, for the sake of landscapes, ecology, and plain NIMBY arguments, is endangering the development of both the wind farms and the necessary power lines.

Also a distributive element becomes important; who is going to pay for the power lines and for the necessary backup by fossil-fueled generation capacity? Indeed, the wind is not always blowing and the sun does not always shine either. The reliability on weather dependent sources of electrical energy on a substantial scale brings in the need for backup capacity, either fossil fuel driven or by means of storage facilities. This requires the development of new market arrangements, as these essentially fixed cost assets will be used only irregularly, while market prices will be unpredictable.

It is obvious that a variety of "right" societal values is being touched upon by green energy developments. The "efficient" market is not providing much of a solution, particularly if the (transaction) costs of observing, measuring, and monetarizing such external and "value dependent" effects are taken into account. Other processes of deliberation may seem more appropriate, but which? And how to achieve such solutions that dynamic efficiency and innovativeness are stimulated, so that "appropriate" technical and institutional solutions will be developed to solve reliability issues over the longer term?

### The Value of Security of Energy Supply

Sustainability, however, was not the only "right" value to be incorporated newly into the established practices of EU market design. From the turn of the century, the economic expansion and the surging energy consumption of the so-called Brasilia, Russia, India, China, South Africa (BRICS) countries began to put pressure on the supply of petroleum and natural gas, which had seen little new investment in exploration and production as expectations of low prices were prevailing. So when post-2000 energy prices began to rise, the question how to expand energy production arose. There were many areas where the international oil and gas industry had no access, like the Middle East and Latin America, where resources were controlled by national state-owned companies. New energy resources could be tapped in technically difficult environments, like the deep sea or the Arctic regions. But the cost and the risk would be huge (see Correljé and van Geuns 2011).

But it was particularly in Russia and the republics of the former Soviet Union that the chances were promising. Indeed, it was in these newly created countries that economic liberalization and the transition away from the communist system provided access to the international energy companies. Creating economic relationships via investments and trade would allow both the EU and these countries to benefit from opening up their economies and industries. The export of the well-known "market design" nicely fitted the EU's internal market paradigm.

In retrospect, this process did not evolve as anticipated. After a decade of radical reform, experiments with privatization, and economic, political, and social turmoil, Russia and other former Soviet Union republics returned from the path of unbridled liberalization. They shifted toward a pattern of resource exploitation, in which foreign involvement would be much more limited and bound by their own national public (and private and elite) interests. Therewith, the prospect of unlimited European access to a huge resource base at highly favorable conditions vanished, as was experienced by a number of European oil and gas companies (Gustafson 2012).

Thereafter, EU-Russia relations began to cool down. They were put under even more pressure post-2006 when conflicts between Russia and Belorussia and Ukraine, as important transit countries for Russian gas and oil, got out of hand and turned into actual supply distortions for parts of central and Southeast Europe. These events were highly effective in putting the notion of security of supply and geopolitics onto the agenda of EU energy policy. Particularly the central European countries that fell victim to the disruptions had just gotten away from their previous political and economic dependence on the Soviet system. The recent developments around the Crimean peninsula and the Ukraine only added to the perception that energy dependence on Russia is a huge problem.

This shows that both the evolution of the global energy markets and the relationship with its eastern neighbors made the European Commission increasingly sensitive for the "value" of security of supply (Zeniewski and Brancucci 2013). Interestingly, the main issue here is not only the exchange value of traded energy as such. The tensions regarding security of supply/demand are created by

the compound impact upon the energy system of: a) the strategic values of geopolitical and international relations as high-level policy, b) the economic exchange values of international trade in energy in both importing and exporting countries, and c) magnified by the strategic and commercial interdependencies, created by rigid, asset-specific, energy transport and production infrastructures.

The consequences of safeguarding the value of "security of supply" for the instrumental values of energy system governance are enormous, however. Specific EU rules were created to provide new pipelines, reverse flow compressors, liquefied natural gas (LNG) import terminals, and strategic storages, to alleviate future supply distortions, and to create a larger diversity of suppliers (CEC 2010). Moreover, to reduce the potential strategic power of Russia, the requirement of unbundling between supply and transport was extended toward the long distance supply pipelines, thus effectively limiting a dominant downstream role for the Russian gas company Gazprom as a seller in the EU market. This approach aims at achieving strategic security of supply, by the creation of regional gas markets of a sufficient size, with a sufficient number of different sellers, even in parts of Europe where such alternative supply routes would be highly expensive (Glachant et al. 2013). Interestingly, the manner in which the European Commission and the regulatory authorities incorporate the value of security of supply in the existing supply system is presented as the creation of a well-functioning pan European gas market, with all theoretical requirements thereof: many suppliers, alternative transport routes, and unbundled vertical trade relationships. Hence, conceptually and rhetorically, it fits with the structural perspective of the subjectivist notion of value and thus with the EU common market paradigm.

Moreover, for many societal interest groups, sustainable entrepreneurs, and public authorities, the measures to secure supply align nicely with those to create a future sustainable energy system. Their argument is that wind, solar, and biofuels would solve both problems at once, if developed on a sufficient scale. The need to achieve supply security, the future high costs of (depleting) fossil energy supply, and the social cost of global warming are invoked to neutralize the argument that sustainable energy is still too expensive; it is an insurance!

However, the infrastructural and organizational pre-requirements for creating such regional markets for green electricity and natural gas are highly dependent on socialized national and EU investments and imply a large reduction of the freedom of contracting and transacting of the actors in these markets. Moreover, the security of supply arrangements strongly draws on the notion of solidarity between the EU member states.

Such developments take place in the context of considerable shifts in the relative prices for the main sources of energy. High oil prices are keeping gas prices relatively high in Europe and Asia. In the USA, oversupply of unconventional gas is driving US gas prices down to the bottom end of the market while pushing coal toward Europe because the EU ETS is dysfunctional. Hence, in the EU gas is priced out of the power sector, and it remains to be seen how a sufficient gas fired backup capacity will be maintained, on the basis of prices generated by the current European power markets.

As a consequence, new forms of financing and tariffs as well as new coordinative institutions will be required to support backup capacity and transport storage facilities. Therewith, a new phase of market creation is to be expected, in which substantial distributional shifts in costs, revenues, and risk are to be "deliberated" between countries, between groups of consumers, and between producers and infrastructure owners and operators. This, most probably, will involve a redistribution of rights of ownership, technical and economic control, and economic exploitation, which will involve a fundamental revision of the current structures and mental maps of the parties involved (see Correljé 2005; Correljé et al. 2014).

## Design for (Moral) Values

In the perspective of the social theory of value, markets are seen as one among the many potential instruments to realize societal values. A well-designed market can be a tool to realize specific (instrumental) values, like an efficient use of assets, under specific prespecified conditions. But if such conditions do not apply, other tools can be considered to be more appropriate to realize other values, like a more equal distribution of income, a sustainable energy production, or more attention for the cultural heritage in the community. Moreover, designing and implementing markets to allocate goods and services are not "value-free" as the subjectivist theory of value suggests. Not only are markets, as discussed, always institutionalized, reflecting specific property and power distributions. Also, as, for instance, Sandel (2012) points out, the use of markets in particular segments or sectors influences the prevailing norms in such parts of society. Therefore, markets are not value-free and cannot be properly analyzed and evaluated within an isolated "subjective" economic discipline.

This also holds for nonmarket institutions. Democratic, participatory coordination mechanisms that have an impact on the norms in society are value-free neither. In other words: which allocation mechanisms are preferable not only depends on their efficiency attributes. It should also depend on the positive or negative impact on the values and norms a society wants to endorse.

From the OIE perspective, it becomes clear that NCE is and cannot be value-free in the sense that NCE would not strive for any other values than safeguarding the individual subjective values of the economic actors.

> Mainstream theory, that is mainly neoclassical theory, is not so value-neutral as its proponents claim. In fact, neoclassical economics (and Austrian economics) is quite outspoken about one important human value operating in economic life, which is eagerly taken up in economic assumptions, concepts and policy advise: the value of freedom, or liberty. (Van Staveren 1999, p. 17)

The conceptualization of the process of institutional change as one in which reality is constituted through action is grounded in the philosophy of American pragmatism, which forms the foundation of OIE (Bush and Tool 2003; Groenewegen 2011). Design for values in the social theory of values is not about

utopian engineering and blueprint thinking. On the contrary, it is about piecemeal engineering and process thinking.

## Conclusion

In this chapter we addressed the question of the "design for values in economics." For that purpose we made a distinction between the subjectivist value theory (NCE and NIE) and the social theory of value (OIE). We argued how design for values in the former schools of economics focuses on designing the right institutions for markets to reveal subjective values, while the latter considers the design of markets as one of the possible tools out of many other nonmarket and hybrid tools to realize social (moral) values.

Essentially, in NCE, the design issue involves the economist in an analysis of the "distance" between the structural characteristics and actors' conduct in a given real world market, like the EU energy market in our example, and the characteristics of the ideal type of any competitive market. This ideal comprises many independently operating suppliers and many active consumers, plentiful information on transactions taking place, free entry and exit for actors, no abuses of market power and market foreclosure by dominant firms, etc. If any problems are identified and the distance is considered too large, the so-called deontological ethical rules of the market are not respected. In that case the economist would have to advise relevant (competition) authorities to take appropriate (legal) action, like reestablishing the right structural characteristics, providing information to market participants, or by correcting firms' anticompetitive behavior, for example, by fines.

An economist working in the NIE paradigm has a more complex job. He/she would be examining the characteristics of relevant transactions in terms of their asset specificity and the prevailing market circumstances in terms of frequency and uncertainty. Taking notice of such insights, he would then evaluate the adequacy of the prevailing market institutions and governance structures. He also would be looking at the occurrence of positive and negative externalities, judging whether a Coasean solution, creating property rights and a market, or a Pigovian tax or subsidy would be preferable, in terms of efficiency.

Thereupon, the NIE economist would advise firms and public authorities how to create the right institutional embedment for that particular market. Difficult questions may arise. Are observed restrictive arrangements between firms really necessary because of the specific characteristics of the transactions? Or do they actually constitute an (illegal) strategy to create market power? And, when certain expected and socially valuable transactions are not taking place, is this a consequence of the preferences of buyers and sellers in the market? Or are there any externalities involved? Or is it legally impossible to establish the right transactional arrangements, including appropriate contracts, forms of oversight, or even public provision? As regards the answers to such questions, the underlying arguments will be

based on norms or conventions that have surfaced in an evolutionary process, in which actors and analysts internalized them as regularities, driven by their efficient self-interest.

The OIE economist looks at design in a dynamic, holistic, and systemic way. The structures in a specific environment are constituted in the interaction among more or less powerful individuals and collectivities and between the actors and structures. Hence, individual and common values emerge and are constituted and designed in interaction. They can be considered right or wrong and are subject to (e) valuation, being judged and deliberated in their specific context of time and place, regarding their consequences for the well-being of the members of society.

A main design issue is about the institutions that facilitate this process of "social construction," in such a way that individuals have access to the necessary information and the relevant arenas for deliberation and decision-making. To participate, actors should be informed, knowledgeable, and aware of their responsibilities. Both markets and nonmarket institutions may enable individuals to reveal their endogenous preferences and values and offer ways to decide about collective values.

In OIE some virtues are more "right" than others. The preference for any mechanism not only depends on its efficiency but also on its positive or negative impact on other values and norms supported by a society. As stated already, design for values according to the social theory of values is not about utopian engineering or blueprint thinking. On the contrary, it is about understanding sociotechnical processes and piecemeal institutional and technical engineering, in a specific context with politically and economically interested stakeholders, among which processes of learning take place that may alter their "belief systems" over time. Interventions may have consequences that are neither sought nor anticipated: two steps forward, one step back.

**Acknowledgment** This research has been financed by a grant of the Energy Delta Gas Research (EDGaR) program. EDGaR is cofinanced by the Northern Netherlands Provinces, the European Fund for Regional Development, the Ministry of Economic Affairs, and the Province of Groningen.

## References

Aoki M (2001) Toward a comparative institutional analysis. MIT press, Cambridge
Aoki M (2007) Endogenizing institutions and institutional changes. J Inst Econ 3(1):1–31
Ayres CE (1944) The theory of economic progress. The University of North Carolina Press, Chapel Hill
Becker GS (1996) Accounting for tastes. Harvard University Press, Cambridge
Boulding K (1969) Economics as a moral science. Am Econ Rev 59(1):1–12
Bromley DW (2006) Sufficient reason. Volitional pragmatism and the meaning of economic institutions. Princeton University Press, Princeton/Oxford
Bush PD (2009) The neoinstitutionalist theory of value. J Econ Issue 43(2):293–307

Bush PD, Tool MR (2003) Foundational concepts for institutionalist policy making. In: Institutional analysis and economic policy. Springer, New York, pp 1–46
CEC (1988) The internal energy market COM(88) 238 Final, 2 May 1988, Commission of the European Communities
CEC (1996) Directive 96/92/EC of the European Parliament and of the Council of 19 December 1996 concerning common rules for the internal market in electricity. Official Journal of the European Union 1997 L 27:20–29
CEC (2003) Directive 2003/54/EC of the European Parliament and of the Council of 26 June 2003, concerning common rules for the internal market in electricity and repealing Directive 96/92/EC. Official Journal of the European Union 2003 L 176:37–55
CEC (2009a) Directive 2009/72/EC concerning common rules for the internal market in electricity and repealing Directive 2003/54/EC. Official Journal of the European Union 2003 L 211:55–93
CEC (2009b) Directive 2009/73/EC concerning common rules for the internal market in natural gas and repealing Directive 2003/55/EC. Official Journal of the European Union 2003 L 211:94–136
CEC (2010) Regulation (EU) No 994/2010 of the European Parliament and of the Council of 20 October 2010 concerning measures to safeguard security of gas supply and repealing Council Directive 2004/67/EC L 259:1–22
Coase RH (1960) The problem of social cost. J Law Econ 3:1–44
Commons JR (1931) Institutional economic. Am Econ Rev 21:648–657
Commons JR (1934) 1961 Institutional economics, vol 1. University of Wisconsin Press, Madison
Correljé A (2005) Dilemmas in network regulation: the Dutch gas industry. In: Künneke R, Groenewegen J, Correljé A (eds) Innovations in liberalized network industries: between private initiatives and public interest. Edward Elgar, Cheltenham, pp 115–150
Correljé A, de Vries L (2008) Hybrid electricity markets: the problem of explaining different patterns of restructuring. In: Fereidoon PS (ed) Competitive electricity markets: design, implementation and performance, Elsevier global energy policy and economics series. Elsevier, Amsterdam/Boston
Correljé A, Van Geuns L (2011) The oil industry: a dynamic Helix. In: Matthias F, Rolf K (eds) International handbook of network industries: the liberalization of infrastructure. Edgar Elgar, Cheltenham, pp 197–214
Correljé A, Groenewegen J, Jaap BJ (2012) The regulated firm in liberalized network industries. In: Dietrich M, Krafft J (eds) Handbook on the economics and theory of the firm. Edward Elgar, Cheltenham, pp 530–544
Correljé A, Groenleer M, Veldman J (2014) Understanding institutional change: the development of institutions for the regulation of natural gas supply systems in the US and the EU. Competition and Regulation in Networked Industries 15(1):2–13
Costa AN, Castro C (2011) Claiming choice for institutional economics. Journal of Economic Issues 45(3):665–684
Denzau A, North D (1994) Shared mental models: ideologies and institutions. Kyklos 47(1):3–31
Dewey J (1922) 1930 Human nature and conduct: an introduction to social psychology. The Modern Library, New York
Dolfsma W (2004) Institutional economics and the formation of preferences: the advent of pop music. Edward Elgar Publishing, Cheltenham
Glachant J-M, Hallack M, Vazquez M, Ruester S, Ascari S (eds) (2013) Building competitive gas markets in the EU: regulation, supply and demand. Edward Elgar, Cheltenham
Greif A (2006) Institutions and the path to the modern economy: lessons from Medieval trade. Cambridge University Press, Cambridge
Groenewegen J (2011) The Bloomington School and American Institutionalism. Good Soc 20(1):15–36

Groenewegen J (2013) A synthesis of neoclassical and institutional economic price theory? In: Dolfsma W, Kesting S (eds) Interdisciplinary economics: Kenneth E. Boulding's engagement in the sciences. Routledge, London & New York, pp 316–331
Groenewegen J, Vromen JJ (1996) A case for theoretical pluralism. In: Groenewegen J (ed) Transaction cost economics and beyond. Kluwer Academic Publishers, Dordrecht/Boston, pp 365–380
Groenewegen J, De Jong M (2008) Assessing the potential of new institutional economics to explain institutional change: the case of road management liberalization in the Nordic countries. J Inst Econ 4(1):51–71
Groenewegen J, van Spithoven A, van der Berg A (2010) Institutional economics; an introduction. Palgrave MacMillan, New York
Gruchy AG (1972) Contemporary economic thought: the contribution of neo institutional economics. Augustus M. Kelley, Clifton
Gustafson T (2012) Wheel of fortune: the battle for oil and power in Russia. The Belknap Press of Harvard University Press, Cambridge/London
Haaland Matláry J (1997) Energy policy in the European Union. Macmillan, Houndmills
Heilbroner RL (1988) The problem of value. In: Heilbroner RL (ed) Behind the Veil of Economics. Norton, New York, pp 104–133
Hodgson GM (2004) Reclaiming habit for institutional economics. J Econ Psycho 25:651–660
Joskow P (2008) Lessons learned from electricity market liberalization. Energy J, Special Issue. The Future of Electricity: Papers in Honor of David Newbery, IAEE
Latsis SJ (1976) A research programme in economics. In: Latsis J (ed) Method and appraisal in economics. Cambridge University Press, Cambridge, pp 1–41
Nooteboom B (2012) Beyond humanism; the flourishing of life self and other. Palgrave Macmilan, Hampshire
Nooteboom B (2013) A pragmatist theory of innovation. In: Melkas H, Harmaakorpi V (eds) Practice-based innovation: insights, applications and policy implications. Springer, Heidelberg, pp 17–27
North DC (2005) Understanding the process of economic change. Princeton University Press, Princeton
Pérez-Arriaga IJ (2013) Challenges in power sector regulation. In: Pérez-Arriaga IJ (ed) Regulation of the power sector. Springer, London, pp 647–678
Pigou AC (1932) The economics of welfare, 4th edn. Macmillan, London, Online at: http://www.econlib.org/library/NPDBooks/Pigou/pgEW.html
Sandel M (2012) What money can't buy: the moral limits of markets. Penguin, London
Spanjer AR (2009) Regulatory intervention on the dynamic European gas market – neoclassical economics or transaction cost economics? Energy Policy 37(8):3250–3258
Spiller PT (2013) Transaction cost regulation. J Econ Behav Org 89:232–242
Thomas S (2003) The seven brothers. Energy Policy 31(5):393–403
Tool MR (1995) Pricing, valuation and systems, essays in neoinstitutional economics. Edward Elgar, Aldershot
Van Staveren I (1999) Caring for economics – an Aristotelian perspective. Eburon, Delft
Van Staveren I (2007) Beyond utilitarianism and deontology: ethics in economics. Rev Polit Econ 19(1):21–35
Veblen TB (1899, ed. 1975) The theory of the leisure class. Augustus M Kelly Academic Publishers, New York
Veblen TB (1904, ed. 1975) The theory of the business enterprise. Augustus M. Kelly, Clifton
Vermaas P, Kroes P, van de Poel I, Franssen M, Houkes W (2011) A philosophy of technology: from technical artefacts to sociotechnical systems. Synth Lect Eng Tech Soc 6(1):1–134
Wilber Charles K, Harrison RS (1978) The methodological basis of institutional economics: pattern model, storytelling, and holism. J Econ Issue XII(1):61–89

Williamson OE (1975) Markets and hierarchies: analysis and antitrust implications. Free Press, New York
Williamson OE (1979) Transaction-cost economics: the governance of contractual relations. J Law Econ 22(2):233–261
Williamson OE (1998) Transaction cost economics: how it works, where it is headed. De Economist 146(January):23–58
Williamson OE (2000) The new institutional economics: taking stock, looking ahead. J Econ Lit XXXVIII:595–613
Zeniewski P, Brancucci C (2013) Framing new threats: the internal security of gas and electricity networks in the European Union. In: Dyer H, Trombetta MJ (eds) International handbook of energy security. Edward Elgar Publishing, Cheltenham, pp 40–70

# Design for Values in Engineering

Ibo van de Poel

## Contents



I. van de Poel (✉)
Department of Ethics and Philosophy of Technology, Delft University of Technology, Delft, The Netherlands
e-mail: i.r.vandepoel@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_25

**Abstract**

Values have probably always played a role in engineering design. However, in current practices and design methods, the attention for values in engineering design tends to be implicit and unsystematic. Establishing Design for Values in engineering would require overcoming this situation. This contribution discusses which values play a role in engineering and engineering design, describes existing methods and experiences with Design for Values in engineering, and explores how values can be integrated into engineering design and existing design methods, in particular quality function deployment (QFD). It identifies four challenges for Design for Values in engineering: (1) discovery of the values to be included in engineering design; (2) translation of these values into engineering characteristics; (3) choice among design options that meet different values to different degrees; and (4) verification of whether a design indeed embodies the intended values.

**Keywords**

Engineering design • Values • Design methods • Design for Values • Design for X • QFD

## Introduction

This chapter focuses on Design for Values in the traditional engineering disciplines, like civil engineering, mechanical engineering, chemical engineering, and electrical engineering. Values such as safety, human welfare, and sustainability obviously play an important role in design in these disciplines. However, there are few engineering design methods that explicitly pay attention to such values. This chapter will discuss the values that are relevant in engineering and how these values may be incorporated in the various phases of the engineering design process and possible methods for Design for Values in engineering.

Philosophers distinguish between what they call value monism and value pluralism. The first is the thesis that there is ultimately only one value in which all other values or value considerations can be expressed. Value pluralism, on the other hand, states that there exists a plurality of values which cannot, at least not in any straightforward way, be reduced to each other. The two viewpoints also surface in engineering. One might believe that ultimately all value considerations in engineering can be reduced to one value; possible candidates for such an overarching value in engineering are social utility, profit, customer satisfaction, or efficiency. The fact that there is such a range of candidates for the overarching value if one subscribes to value monism already seems to erode the credibility of the thesis that there is ultimately only one value as value monism posits. In this contribution, I will assume value pluralism, i.e., I will assume that a range of values is important in engineering, which cannot easily be reduced to each other.

Another distinction that philosophers often make is that between final (or intrinsic) values and instrumental values. The first are values that are strived for their own sake, while the latter are strived for the sake of other values. The suggestion that comes with this distinction is that final values are more important than instrumental values and there is indeed some truth to this suggestion.

Distinguishing between instrumental and final values does not yet tell us which values are instrumental and which are final. There may be considerable disagreement about this issue. Frankena (1973) lists 18 final values on the basis of a reading of the philosophical literature. This includes moral values like happiness, health, and morally good dispositions, but also nonmoral values like aesthetic experience and truth. In distinguishing between final and instrumental value, I will by and large follow Frankena's list. So I will assume that values like human well-being, justice, safety, health, and sustainability that play a role in engineering are final values, while values like economic profit, efficiency, reliability, and maintainability, which are obviously also important in engineering, are instrumental values.

I start this contribution with a brief description of engineering as a profession and the values that have been articulated as important in engineering generally and more specifically in engineering design. I will then look a bit deeper into the engineering design process, the different phases that might be distinguished in engineering design, and how value considerations may play a role in these. Next, I turn to design approaches to Design for Values in engineering. After discussing challenges and future work, I end with a brief conclusion.

## Values in Engineering

While there has been quite some attention for the relation between values and technology, less attention has been paid to the role of values in engineering. I will understand engineering here as an activity that is aimed at understanding, creating, improving, and maintaining certain technologies (Van de Poel 2010). Values in engineering originate in part from the values that are to be realized by technology. Such values are, for example, incorporated in the engineering design process (Van de Poel 2009). Think of values like safety, sustainability, and human well-being. Values can, however, also emerge in engineering because it is a professional practice (Davis 1998; Pritchard 2009). Examples are values like integrity, honesty, loyalty, and independence. For determining the values of engineering, I start with discussing engineering as a profession and briefly consider the history of engineering as a profession. Then I discuss professional codes for engineers and the values they explicate. Finally I give an overview of some of the main instrumental and final values in engineering design.

## Engineering as a Profession

Especially in the American literature on engineering ethics, engineering is often portrayed as a profession (Davis 1998; Harris et al. 2013). Michael Davis defines a profession as "a number of individuals in the same occupation voluntarily organized to earn a living by openly serving a certain moral ideal in a morally-permissible way beyond what law, market, and morality would otherwise require" (Davis 1998, p. 417). He believes that engineering, at least in most countries today, is a profession according to this definition because most engineering societies have committed themselves (voluntary) to hold paramount the safety, health, and welfare of the public. They have done so by formulating professional codes that I will discuss below. Before I do so, I briefly sketch the historical development of engineering as a profession for pointing out that the values held in engineering can differ among different engineering professions.

Historically, engineering was in many countries closely tied to the military and to a number of nonprofessional occupations like architect-inventors (such as Leonardo da Vinci), instrument-makers, land surveyors, millwrights, masons, and carpenters (Calhoun 1960, pp. 5–6). It was with the emergence of civil engineering that engineering emancipated itself from the military and these occupations and became a more independent professional activity. Civil engineering was established as a profession in the late eighteenth century in France (Abbott 1988, p. 92). In the USA, civil engineering became a profession in the second half of the nineteenth century (Calhoun 1960). The pattern of development was different from country to country: whereas in France and the Netherlands military engineers were the main predecessors, they were millwrights and instrument-makers in Great Britain (Calhoun 1960, p. 7; Lintsen 1985, pp. 16–22).

In the course of time, new engineering professions have emerged such as mining engineering, mechanical engineering, electrical engineering, and chemical engineering (Calvert 1967; Reynolds 1983). Sometimes the development of new engineering professions was due to the development of new technologies or to inventions, for example, in the case of nuclear engineering. In many cases, new professions had to emancipate themselves from already existing professions. Mining and mechanical engineering emancipated themselves from civil engineering in the late nineteenth century. Chemical engineering emancipated itself from mechanical engineering and chemistry.

What is interesting from the point of values in engineering is that the various engineering professions also seem to have somewhat different value sets. So physical safety and health are main values in, for example, chemical engineering, but they are, in general, less important in, for example, software engineering. Whereas in software engineering, privacy is a main value, it is less important, or sometimes even irrelevant, in chemical engineering. Despite the differences between engineering disciplines, there is also much communality to the values in engineering. In the remainder of this section, I will mainly focus on the values that are shared among most engineering professions. Nevertheless, it is important

to keep in mind that in addition to these more general values, there are also values that are more specific to particular engineering professions.

## Professional Codes in Engineering

Professional codes formulated by engineering societies are often seen as a main expression of the values that are typical for engineering.[1] Such codes are often aspirational: they express the main values of engineering in rather general and abstract terms without the aim of giving detailed advice about how to behave in concrete situations or the aim of regulating professional behavior in detail.

Historically, the development of professional codes for engineers began in England in 1771 with the code of the *Smeatonian Society*. More influential for the current professional codes for engineers was the formulation of a range of professional codes for different engineering professions like civil, mechanical, and electrical engineering in the first decade of the twentieth century in the USA. The early codes comprised rules for engineers that chiefly pertained to etiquette. The professional code regulated people's entry into the profession and the behavior of members toward each other and in relation to employers and clients. While the early codes did not address broader social issues raised by engineering, this changed after the Second World War. The duty of the engineer to serve the public interest was especially stressed in the new professional codes. Organizations like the National Society of Professional Engineers (NSPE), the American Society of Civil Engineers (ASCE), and the American Society of Mechanical Engineering (ASME) formulated professional codes stating that engineers "should hold paramount the safety, health and welfare of the public." Similar values are emphasized in the universal statement regarding the conduct of professional engineers that was issued by the European Federation of National Engineering Associations (FEANI).

Professional codes for engineers thus express the core values of the engineering profession. Most modern professional engineering codes relate to three domains: (1) conducting a profession with integrity and in a competent way, (2) obligations toward clients and employers, and (3) responsibility towards the public and society.

### Integrity and Competent Professional Practice

All professional codes include the obligation to practice one's profession with integrity and honesty and in a competent way. The traditional core of all professional codes thus stresses such values as honesty, faithfulness, truthfulness, integrity, and competence.

[1]This section draws on Van de Poel and Royakkers (2011), Chap. 2.

### Obligations Toward Clients and Employers

Obligations toward clients and employers are mentioned in most professional codes. In many cases, it is stipulated that engineers should serve the interests of their clients and employers and that they must keep secret the confidential information passed on by clients or employers. Values that are stressed in this respect include loyalty, confidence, trustworthiness, and fair play.

### Social Responsibility and Obligations Toward the Public

Virtually all professional codes in one way or another emphasize the social responsibility of engineers. Values that are often mentioned include safety, health, the environment, sustainable development, and the welfare of the public.

## Instrumental and Final Values in Engineering Design

I will now further explore and define some of the main values that play a role in engineering. In doing so my focus will be on values that engineers somehow try to incorporate in the technologies they research, develop, and design. The reason for this focus is that this contribution is about Design for Values in engineering, a focus that is narrower than just values in engineering. For this reason, I will not pay attention to such values as integrity, honesty, impartiality, expertise, loyalty, and rationality, which are all more related to engineering as a process (or an activity, if one wishes) rather than being values for which engineers design – develop and research – technologies. Below, I will follow the distinction between instrumental and final values that I introduced in the introduction.

### Instrumental Values

We might distinguish the following instrumental values that are relevant to engineering design:

- *Effectiveness*, which may be defined as the degree to which an artifact fulfills its function.
- *Efficiency*, which could be defined as the ratio between the degree to which an artifact fulfills its function and the effort required to achieve that effect. Efficiency in the modern sense is usually construed as an output/input ratio (Alexander 2009).
- *Reliability*, which might be understood as "the ability of a product to perform its function adequately over a period of time without failing" (cf. Kuo et al. 2001, p. 252).
- *Robustness*, which may be defined as the "ability of a product to perform its function adequately in new or unforeseen circumstances" (cf. Vermaas et al. 2011, p. 113).
- *Maintainability*, which might be understood as "the probability that a failed system can be repaired in a specific interval of downtime against reasonable cost" (cf. Kuo et al. 2001, p. 251).

- *Compatibility*, which might be understood as the ability of a product to adequately perform its function in conjunction with other apparatus and infrastructure.
- *Quality*, which might be understood in a variety of ways. Sometimes it is used to refer to such values as reliability, robustness, and compatibility. It is also used in the sense of "robust in meeting the requirements (within certain acceptable limits) despite variations in the production process" (cf. Holt and Barnes 2010, p. 125). It might also be understood in terms of "meeting or even exceeding user requirements" or in terms of "user satisfaction."

### Final Values

In addition to these instrumental values, the following final values are relevant to engineering design and are often mentioned as paramount in engineering (in, e.g., professional codes):

- *Safety*, which is sometimes defined as the absence of risk and hazards. However, risk reduction is not always feasible or desirable. Safety is therefore maybe best understood in terms of "acceptable risk." The ethical literature on risk has established that the moral acceptability of risks does not only depend on their magnitude but also on considerations like voluntariness, the balance and distribution of benefits and risks, and the availability of alternatives (Asveld and Roeser 2009; Hansson 2003, 2009; Shrader-Frechette 1991). So conceived, safety refers to the situation in which the risks have been reduced in as far that is reasonably feasible and desirable.
- *Health*, which is defined by the World Health Organization (WHO) as "state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (World Health Organization 2006). In engineering, the focus is usually on avoiding negative influences on human health. It is not obvious that there is a requirement for engineering to contribute positively to human health, with the exception perhaps of some specific domains like health technologies.
- *Human well-being*. This value is being referred to under a number of headings like human welfare, happiness, quality of life, human flourishing, and good life. I will here use the term "human well-being" to refer to the value that is at stake in all these cases. Well-being does not just refer to feeling well here and now but it tells something about how someone's life is going for that person.
- *Sustainability*. Although environmental values play a role in engineering for quite some time, in the last decade this has been increasingly understood in terms of the broader value of sustainability. The most influential definition of sustainable development has been provided by the Brundtland Commission: "Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs" (WCED 1987).

Other final values are relevant for engineering as well. Some of these final values are generally relevant for engineering. Examples are justice and democracy, and inclusiveness (Sclove 1995; Clarkson 2003; Erlandson 2008). In addition to such more general final values, one might distinguish final values that are more domain-specific. A typical example is aesthetics in architecture or privacy in ICT.

## Incorporating Values in the Engineering Design Process

Engineering design is the process by which certain functions are translated into a blueprint for an artifact, a system, or a service that can fulfill these said functions. In traditional design methods, the engineering design process is usually depicted as a systematic process in which use is made of technical and scientific knowledge, but in which creativity and decision-making also play major roles. (For a discussion of some more recent design approaches and their relevance for Design for Values, see the chapters on "► Design Methods in Design for Values" and "► Participatory Design and Design for Values").

Design methods usually divide the design process in different stages between which iterations are possible (Pahl et al. 2007; Hubka 1982; Roozenburg and Cross 1991; Eekels and Roozenburg 1991; Ullman 1997; Cross 2008).[2] Although the exact stages are different from design method to design method, many of them contain a number of basic activities like analysis (of the design problem), synthesis (of possible design solutions), evaluation (of the possible solutions in the light of the problem), and choice (of one design solution). Additional steps that are often mentioned include simulation, embodiment design, and prototype testing. Simulation refers to making predictions about how possible design solutions (concept designs) will behave, a step that might involve calculation, modeling, testing, etc. Embodiment design is the phase that follows after one design solution has been chosen and has to be further detailed, finally resulting in design drawings and technical specification on basis of which the design can be built or produced. Prototype testing refers to the testing of prototypes of the system, possibly resulting in new insights and reiterations of the design process. Figure 1 depicts the basic stages of the design process.[3]

In all the mentioned phases of the design process, values play a role but the role they play is (quite) different in the various phases as I explicate below.

[2]Not all design methods conceptualize the design process as a linear process. Most methods contain possibilities for iteration. Moreover, especially design models from architecture stress that the design problem cannot be formulated completely independent from possible solutions (Roozenburg and Cross 1991, p. 188).

[3]The figure is largely based on Eekels and Roozenburg (1991).

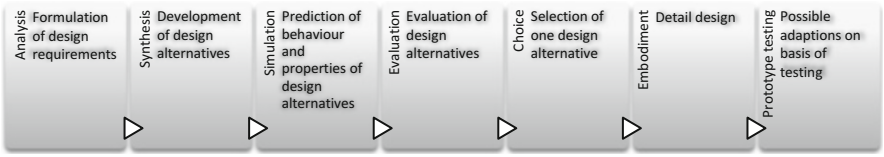



**Fig. 1** Phases of the design process

## Analysis

During the analysis phase, the designer or the design team conceptualizes the design problem. This stage results in a certain formulation of the design problem and of certain design requirements that a good or acceptable solution has to meet. Findings in later stages can sometimes result in the revision of the problem formulation or the design requirements.

Values play a role in this stage in several ways. First, they will influence how the problem is conceived and framed. During the analysis stage, the designers also might make an inventory of relevant values based on, for example, the design brief, professional codes, and legislation or by inquiring the stakeholders. Values are in this stage also relevant as a source of design requirements (Van de Poel 2013).

## Synthesis

In this phase the designer or design team thinks out potential solutions to a design problem. The focus is on an integral approach to the design problem. The designer does not try to realize each design requirement independently but works on a combination of design requirements and searches for a total concept that can bring about this combination. Creativity is important in this phase, especially for thinking out new solutions that might meet seemingly conflicting requirements or values. In this stage, the values that have been identified in the analysis stage and have been translated in that stage in design requirements are embodied in various conceptual design solutions.

## Simulation

The concept designs are checked in the simulation stage to see whether they meet the design requirements. This takes place in a number of ways, e.g., through calculations, modeling, and computer simulations. Modeling plays an important part in this phase, and the models that one develops or uses should be appropriate to predict the effects of the conceptual design solutions on the various relevant value dimensions that have been identified in the analysis stage (see chapter on "► Modeling for Design for Values"). It might also be that in this phase,

unexpected consequences of the conceptual design may lead to the introduction of new values that should be considered in the design problem.

## Evaluation

During the evaluation phase, the outcomes of the simulation of different design solutions are evaluated. They may be evaluated in terms of meeting the functional requirements or costs, but usually engineering values – instrumental as well as final – play a major part in this phase. Some of the values like efficiency may be readily measurable, but especially final moral values often will first need to be operationalized before their attainment can be measured and evaluated (see the chapter "► Design for Values and the Definition, Specification, and Operationalization of Values").

## Choice

In this stage, a choice is made which concept design is to be detailed further. This choice is based on the outcomes of the evaluation of the various concept designs. Values play a role as decision criteria in this phase. Often it will not be possible to choose a design that meets all design requirements and that meets all values to a maximum degree. So often a choice need to be made under conflicting values (see chapter on "► Conflicting Values in Design for Values").

## Embodiment

In the embodiment phase, one design is further detailed. In this phase, the relevant values are further translated and embodied in engineering characteristics of the design. In this phase, also choices are made with respect to, for example, the materials of which the design will be made and the production methods. This may introduce additional values, for example, related to the use of scarce materials or to labor circumstances of the producers of the product.

## Prototype Testing

In this phase, a prototype of the design is tried out, which may lead to iterations in the design process. Prototype testing may be particularly relevant for verifying whether the designed system indeed embodies the intended values. But prototype testing may also point out unexpected consequences or value dimensions of the design. It may turn out that the design has unexpectedly certain health consequences, which means that an iteration of the design process is required taking the value of health into account.

**Table 1** The roles of values in different stages of the design process

| Stage | How values play a role | Activity |
|---|---|---|
| **Analysis** | In framing of design problem<br>Discovery of relevant values<br>Translation in design requirements | Discovery<br>Translation |
| **Synthesis** | Embodiment of values in various concept designs | Translation |
| **Simulation** | As dimensions that should be included in modeling and simulation<br>Potential discovery of new relevant values | Verification<br>Discovery |
| **Evaluation** | As evaluation criteria<br>Need for operationalization of values | Verification |
| **Choice** | As choice criteria<br>Need for choice under conflicting values | Choice |
| **Embodiment** | Embodiment of values in detail design<br>New value may be relevant for detailed choices | Translation<br>Discovery |
| **Prototype testing** | Verification of values<br>Potential discovery of new relevant values | Verification<br>Discovery |

I have summarized the different roles that values play in the different phases of the design process in Table 1. In the final column, I have related the various phases to the three steps or activities that are distinguished in a design method for Design for Values that has been proposed by Flanagan et al. (2008). Their method consists of three steps or activities:

- *Discovery*. This activity will result in a list of values that are relevant for the design project.
- *Translation*. Translation is "the activity of embodying or expressing . . . values in system design" (Flanagan et al. 2008 p. 338).
- *Verification*. This is assessing, e.g., trough simulation, tests, or user questionnaires, whether the design indeed has implemented the values that were aimed at.

The table suggests that these steps or activities can indeed be associated with the various stages distinguished in traditional engineering design methods. It also suggests, however, certain additions to their method. First, it shows that particularly discovery is not an activity that is, or should be, restricted to the first design stage but is a continuously ongoing activity.

Second, the engineering design methods contain a step that is missing in the method of Flanagan et al: choice. It must be said, however, that there are also design methods in engineering that do not distinguish a separate choice stage. Also in practice, engineering designers are reported to often follow a single-concept strategy method (e.g., Stauffer et al. 1987; Ullman et al. 1987; Bucciarelli 1994; Henderson 1991; Visser 1990, 2009; Stauffer and Ullman 1988). Designers often quickly move to a possible solution, which may be based on their experience or

existing products. They will then try to further develop and improve that solution till it meets the design requirements. When it turns out to be impossible to meet the requirements, they may abandon their original solution concept and try a new one.

However, even if a single concept strategy is followed, the designer will go through different iterations of the steps discovery, translation, and verification that are distinguished by Flanagan et al. Moreover, they make at least implicit choices about whether the design developed is good enough, or to further develop it, or to consider another design option.

The upshot of the discussion then is that discovery, translation, verification, and choice are best seen as activities that are required for Design for Values even if they may not be strictly associated with certain stages of the design process.

## Approaches to Design for Values in Engineering

The idea of incorporating values in design is not new. It might be argued that the practice of incorporating values in the design of technology is as old as technology itself or at least as old as engineering design as a specific activity in the development of technology. In recent decades, a number of authors have developed approaches for what might be called Design for Values.

I use the term Design for Values here for a number of approaches which come under different headings like Value-Sensitive Design (VSD), Values at Play, and Value-Conscious Design (Flanagan et al. 2008; Friedman 1996; Friedman and Kahn 2003; Friedman et al. 2006; Manders-Huits 2011, see also the chapter on "► Value Sensitive Design: Applications, Adaptations, and Critiques"; Cummings 2006; Albrechtslund 2007; Van den Hoven 2005). The general thrust of these approaches is the integration of values of ethical importance in a systematic way within engineering design. Design for Values has been particularly articulated in the domain of information systems and software (see chapter on "► Design for Values in ICT"). However, the idea is more generally applicable; it applies to all kinds of engineering design and in fact also to many kinds of nonengineering design like architectural design, policy design, or institutional design.

Design for Values is related to a number of other approaches in engineering and technological development. One family of approaches is Technology Assessment (TA) (Grunwald 2009). Traditional TA aims at predicting the social consequences of technological development. Although traditionally TA was not aimed at influencing design, most of the recent TA approaches try to incorporate social concerns and values at the design stage. This includes approaches such as Constructive Technology Assessment (CTA), Interactive Technology Assessment (ITA), and Real-Time Technology Assessment (Reuzel et al. 2001; Rip et al. 1995; Schot and Rip 1997; Grin and van der Graaf 1996; Guston and Sarewitz 2002; Grin and Hoppe 1995; see also the chapters on "► Technology Assessment and Design for Values" and "► Design for Values in Healthcare Technology").

Although the general motivation of such approaches is the same as Design for Values, there is a difference in emphasis. TA approaches focus more on social concerns than on values. The approaches are moreover more procedural and sociological in nature, focus less on moral issues, and are less philosophically informed.

A second family of approaches is those that are critical of current technological development and that propose alternative approaches to technological development or engineering design. Some of these approaches are activist in nature. Such approaches can be found in the philosophy of technology (Winner 1986; Feenberg 1995; Sclove 1995) but also in the literature on design (Papanek 1984, 1995; Whiteley 1993). Sometimes authors merely criticize current technological development without offering an alternative or only stating alternatives in very general and abstract terms. Others have proposed more concrete alternative approaches under such names as socially responsible design, ethical design, design for the real world, design for society, and feminist design (Papanek 1984; Tatum 2004; Feng 2000; Whiteley 1993; Nieusma 2004). In many cases these alternative approaches either stay very general or they rather quickly move to practical guidelines, tools, and methods without considering the values at stake in any depth.

A third family of related approaches is based on ideas of concurrent engineering and "design for X" (DFX) that have been articulated in the engineering literature. Concurrent engineering is an approach to engineering in which downstream considerations, such as production, use, and maintenance, are integrated into upstream decisions in engineering design and development. In DFX approaches, X can stand for a certain value or for a life phase. $DFX_{lifephase}$ approaches include, for example, design for manufacture, design for assembly, design for disassembly, design for maintenance, design for recycling, and design for supply chain (Boothroyd et al. 2011; Holt and Barnes 2010; Bogue 2007, 2012; Kuo et al. 2001; Gaustad et al. 2010; Slater 2000; Manohar and Ishii 2009). $DFX_{value}$ approaches for instrumental values include design for quality (like quality function deployment or QFD), design for reliability, and design for robustness (Raheja and Gullo 2012; King 1989; Hauser and Clausing 1988; Park and Antony 2008; Ireson et al. 1996; Cheng 2009; Akao 1990). Also $DFX_{value}$ approaches for final values have been developed, sometimes within the engineering literature, sometimes in other areas. This includes affective design and emphatic design (Jordan 2003; Koskinen et al. 2003), inclusive and universal design (Keates and Clarkson 2003; Clarkson 2003; Imrie and Hall 2001; Erlandson 2008; Preiser and Ostroff 2001; see also the chapter on "► Design for the Value of Inclusiveness"), safety engineering and safe design (Hansson 2007; see also the chapter on "► Design for the Value of Safety"), ecological design and design for sustainability (Bhamra and Lofthouse 2007; Birkeland 2002; Van der Ryn and Cowan 2007; see also the chapter on "► Design for the Value of Sustainability"), and design for capabilities and design for human well-being (Van de Poel 2012; Desmet and Pohlmeyer 2013; Oosterlaken 2009; see also the chapter on "► Design for the Value of Human Well-Being").

## Toward an Integrated Approach: Quality Function Deployment (QFD)

Of the various mentioned approaches, the VSD approach and DFX approaches are probably most directly relevant and applicable in engineering design. Most DFX approaches, however, focus on specific values. As Holt and Barnes (2010) argue in their overview article on DFX approaches, there is a need for an integrated DFX approach that combines different values, that is able to deal with trade-offs between values, and that offers decision support for such trade-offs decisions. The VSD approach, which was mainly developed in ICT, offers an integrated approach that could also be applied to engineering design. However, it does not explicitly address trade-offs and does not offer decision support for such decisions. It might therefore be interesting to look at an approach that can deal with trade-offs although it does not yet focus on a range of values: quality function deployment or QFD. The value that is central in QFD is user value or, more precisely, customer satisfaction.[4] However, as I will show below, the QFD approach can be adapted to a more integrated approach that includes a range of values in engineering design.

### The Traditional QFD Approach

Quality function deployment was originally developed in Japan in the late 1960s[5]. It is now widely used not only in Japan but also in Europe and the USA. The introduction of QFD, and other quality methods, in especially the USA was a response to the growing success of the Japanese industry during the 1970s. QFD was seen as an important tool to improve quality, to reduce development and other preproduction costs, to increase organization capabilities, and – all in all – to make the American industry more competitive. Apart from such business goals, QFD has been heralded as a means for the development of products that better fulfill users' needs.

A main goal of QFD is to translate customer demands into engineering characteristics. By systematically and quantitatively employing the relationship between customer demands and engineering characteristics, those engineering characteristics that are most promising for improving customer satisfaction can be selected. In this way, QFD leads to a more systematic attention for customer demands in the design and development process.

A central element in the QFD method is the so-called House of Quality (Fig. 2). This House of Quality relates customer demands to engineering characteristics. The idea is that in this way, the desires of customers can be translated into (numerical) target values for the engineering characteristics and into priorities for improving certain engineering characteristics.

[4]Customer satisfaction may, depending on one's theory of human well-being, be seen as a value that contributes to the value of human well-being (see chapter on "▶ Design for the Value of Human Well-Being").

[5]This section draws on Van de Poel (2007).

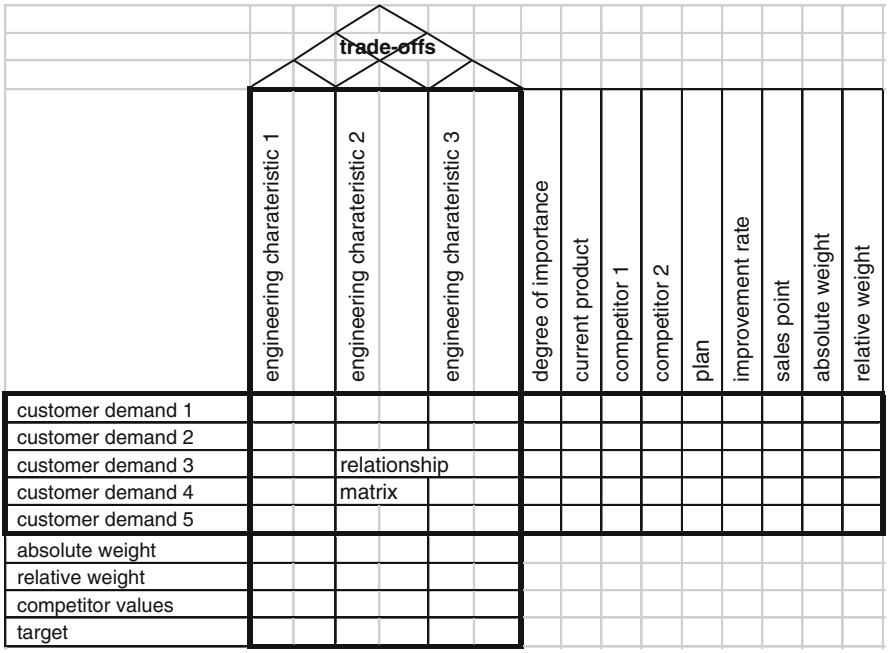



**Fig. 2** House of Quality

Filling in the House of Quality starts with listing the customer demands in the rows in the central part of the house. Subsequently, the degree of importance of the customer demands is filled in. The score of the own existing product and that of competitors with respects to the customer demands are then listed, usually on an integer scale from 1 to 5. On the basis of this competitive benchmarking and strategic considerations, the company plan for each customer demand is chosen, again on an integer scale from 1 to 5. The rate of improvement is calculated by dividing the company plan by the current company score.[6] Next, sales points are set for customer demands that are expected to influence sales more than average. Sales points usually take the values 1.5, 1.2, or 1. The absolute weight of the customer demands is calculated by multiplying the degree of importance with the rate of improvement and the sales point (Akao 1990; King 1989).

The next step is relating the customer demands to the engineering characteristics. To achieve this, first the engineering characteristics are listed in the columns in the central part of the House of Quality. Next the relationship matrix is filled in, using symbols like ◎ (strong correlation), ○ (moderate correlation), and

[6]It might be argued, however, that since both the company plan and the current company score are measured on an ordinal scale (expressed in the integers 1,2,3,4,5), this division is not allowed because ordinal scales do not allow for this arithmetical operation.

△ (weak correlation), which are presumed to correspond with numerical values like, for example, 9, 3, and 1. On the basis of the weighted customer demand and the relationship matrix, the relative importance of the engineering characteristics is calculated. The numerical values of the engineering characteristics for the current product and those of competitors are listed, and targets for the engineering characteristics may be set. As a final step, the trade-offs between the engineering characteristics are listed in the roof of the House of Quality. Usually five types of relations between engineering characteristics are used: strong positive, weak positive, no relation, weak negative, and strong negative.

The House of Quality thus gives insight in the relative importance of the engineering characteristics based on the customer demands. This relative importance may be used to set priorities in further design and development efforts or to select among different conceptual designs. By making a number of further charts, the relative importance of customer demands or engineering characteristics can also be translated into relative weights for certain functions, mechanisms, parts, process steps, and failure mechanisms and in setting priorities for these and for cost reduction (Akao 1990; King 1989).

Also (numerical) target values for the engineering characteristics can be determined. Apart from the relative weights of the engineering characteristics, estimates about what is technically feasible against what costs and efforts, and strategic considerations at the company level do play a role in setting targets. In the initial method, setting targets was left to the discretion of the engineers on the basis of the filled in House of Quality.

### QFD as an Integral Approach to Design for Values

The focus in QFD is on customer demands and ultimately on the value of customer satisfaction. Nevertheless, it is possible to include a range of other values into QFD. Indeed, QFD theorists and practitioners have already tried to incorporate additional kinds of considerations in QFD. In the original QFD method proposed by Akao, for example, the relative importance of the customer demands is not just based on what customers want, but also on the basis of considerations of the producers, like what is the company good at compared to competitors and which customer demands will probably raise sales (Akao 1990). With respect to regulatory requirements, some authors have suggested that these could be treated as customer demands in QFD (Govers 1996).

In line with, and expanding these suggestions, I think there are three main possibilities for incorporating additional values, in addition to customer satisfaction, in QFD:

- The values could be treated as or translated into demands that are treated in the QFD similarly to the customer demands. An advantage of this method is that the relevant values are met as good as possible and desirable in the light of the other relevant demands and the technical possibilities. A potential disadvantage of this approach is that it does not define a minimum level for the values below which products are not acceptable.

- The values could be treated as or translated into minimal requirements to be met by each alternative. The advantage of doing this is that one is sure that the relevant values are respected by any alternative. A possible disadvantage is that only the minimal is met while it may be desirable to do more than the minimal.
- Values could also be used to "correct" the relative importance of the (customer) demands. This is the approach chosen with respect to considerations on behalf of the producer in the original QFD approach. An advantage of this approach is that values are treated as overall considerations rather than as demands besides the other demands (as in the other options). This may be more appropriate, at least for some value considerations like the concern for safety. A disadvantage might be that this way of "correcting" the outcomes of the QFD matrix might be rather arbitrary because we lack an adequate way, let alone methodology, to carry out such corrections.

It seems thus possible to employ QFD as an integral approach to Design for Values. Main advantages of the QFD approach are (1) that it supports the *translation* of values into engineering characteristics and (2) that it helps to trace possible *trade-offs* (in the roof of the House of Quality). Does the approach also offer decision support for making choices in design? It certainly aims to do so. In fact, a range more or less sophisticated quantitative approaches to QFD have been developed that offer decision support. Most of these approaches do so by aiming at a maximization of customer satisfaction. As I have pointed out elsewhere (Van de Poel 2007), such QFD approaches are beset with methodological problems which make it questionable whether they can indeed maximize customer satisfaction.[7] When QFD is extended to include other values in addition to user values, these methodological problems will likely increase (see also the chapter on "► Conflicting Values in Design for Values"). It is therefore doubtful whether QFD can offer decision support for trade-offs between values in Design for Values.

## Challenges and Future Work

In section "Incorporating Values in the Engineering Design Process," I distinguished four activities in engineering design that are crucial for Design for Values: (1) discovery of values, (2) translation of values into design requirements and engineering characteristics, (3) choice: design support for trade-offs between values, and (4) verification of values. In the ideal case, approaches or methods for Design for Values should support all these activities. As we have seen QFD can offer support for activity 2 (translation) and some support for 3 (choice). VSD offers support for activity 1 (discovery) and some support for activity 2 (translation)

[7]However, a more qualitative approach to QFD might be possible that at least is likely to increase customer satisfaction compared with the current situation.

and activity 4 (verification) but hardly for activity 3 (choice). Although various DFX offer support for these four activities for specific values, they usually do not in an integral way that does justice to a range of values.

As far as I know, there is currently no method or approach that supports all four mentioned activities and does so in an integral way. One challenge for the future then is to develop such a method or approach. This might not require a completely new method or approach, as it can also build on, extend, or combine existing methods or approaches. In addition to this more general challenge, there are a number of more specific challenges related to these four activities. It could be argued that these more specific challenges should first be tackled before it is useful to develop a more general Design for Values approach.

## Discovery

Both in engineering design (see, e.g., Pahl et al. 2007) and in VSD (see chapter on "► Value Sensitive Design: Applications, Adaptations, and Critiques"), methods have been developed that are useful for the discovery and the elicitation of values. Especially in VSD, more general social scientific methods are used for this purpose like interviews, surveys, scenarios, participant observation, and ethnographic research. However, the activity of value discovery in design requires more than just identifying potentially important values. Designers also need to answer the *normative* question what values are worth pursuing in design. This normative question in turns raises a number of more foundational and metaphysical philosophical questions about values. Is value subjective or objective? Should we distinguish between final and instrumental values? Are there universal values or are value relative to culture and place? Manders-Huits (2011) argues that VSD lacks a normative criterion to decide what values should be included in the design and that it should therefore be supplemented by a normative moral theory of values.

## Translation

The translation of values into the materiality of the object designed might be broken down into at least two steps: (1) the specification of general values in terms of design requirements and (2) the translation of such requirements into engineering characteristics. For the first step use may be made of methods and approaches developed in requirements engineering (e.g., Hull et al. 2005; Young 2003; Grady 1993) or in decision theory (e.g., Keeney 1992). Building on these, Van de Poel (2013) proposes an approach for translating values into design requirements. Also the chapter "► Design for Values and the Definition, Specification, and Operationalization of Values" is relevant.

The second translation, from requirements to engineering characteristics, is made, for example, in QFD as we have seen, although QFD does not offer specific

guidelines or suggestions on how to make this translation. At the background here is the question how QFD is able to translate descriptions in the functional or value domain in to descriptions in the structural or physical domain; some of these issues are discussed in De Vries (2009). The more fundamental philosophical issue here is the idea that designed technical artifacts have a dual nature, they are both intentional social objects and have a technical structure that can be understood in terms of natural science, and the question is how we translate requirements in the intentional or value domain into characteristics in the technical, natural domain (e.g., Kroes 2010).

## Choice

There is a plethora of methods for making choices in engineering design (e.g., Pahl et al. 2007; Cross 2008; Pugh 1991; Dym et al. 2014; Akao 1990) and also more generally in the area of decision theory (e.g., Keeney and Raiffa 1993) for dealing with cases of conflicting criteria or values in design. As has been shown by Franssen (2005), many of these methods run into Arrow's Theorem, a well-known impossibility theory from decision theory (or they do not actually offer decision support but only reconstruct decisions).

When the choice to be made in engineering design is not just understood as a multi-criteria choice problem, but also as a problem in terms of multiple values, this raises additional philosophical issues related to what philosophers have called value incommensurability (Raz 1986; Chang 1997). Two or more values are incommensurable if they cannot be measured on the same scale. Incommensurability may arise from the fact that it is impossible or at least inappropriate to cancel out losses in one value domain by benefits in another value domain. Value incommensurability raises fundamental philosophical questions about choices under value conflict in design and the rationality of such choices. Some of the relevant issues are discussed in Van de Poel (2009) and in the chapter "▶ Conflicting Values in Design for Values."

## Verification

Verification in Design for Values is, as far as I am able to tell, still largely an unresolved issue on which hardly work has been done. There are standard (social science) methods for evaluating whether a design *in use* meets or respects certain values according to the *current users*. It would, however, be most helpful to have methods to evaluate or validate proposed designs with respect to their incorporated values already in the design phase before they are actually used. Methods may be available for specific value, like safety, but no general approaches seem to have been established.

The underlying philosophical issue is whether values can be embedded in designed technical artifacts, and if so what it exactly means to say that a design embeds or embodies certain values, an issue that obviously needs to be clarified (or at least assumed) if verification methods for Design for Values are to be developed.

The issues of the value-ladenness of technology and how it is exactly to be understood is hotly debated; different positions are discussed in Kroes and Verbeek (2014).

## Conclusions

Values have probably always played a role in engineering design. However, most current engineering design practices and methods are characterized by an implicit and unsystematic attention for values in the design process. In as far as values have received explicit and systematic attention, the focus has often been on more instrumental values such as effectiveness, efficiency, reliability, and customer satisfaction, but there are also some design methods or approaches that explicitly address final values in engineering design. What is still missing yet is an integral approach to Design for Values that offers support to four key activities in Design Values, i.e., (1) discovery of values, (2) translation of these values into engineering characteristics, (3) choice about conflicting values and trade-offs, and (4) verification of values in the designed product.

It remains to be seen whether Design for Values in engineering requires the development of completely new methods. It might be possible and more effective to build on existing methods and approaches such as QFD. More important than developing new design methods as such is the addressing of the four challenges for Design for Values in engineering that I have identified. These are the challenges: (1) How to decide what values should be incorporated from a normative point of view in a design? (2) How to make the translation from the intentional, functional domain to the structural, natural domain for values? (3) How to deal with conflicting values in design? (4) How to verify or validate whether a design indeed embodies or represents the values for which it has been designed?

## Cross-References


- ▶ Conflicting Values in Design for Values
- ▶ Design for the Value of Human Well-Being
- ▶ Design for the Value of Inclusiveness
- ▶ Design for the Value of Safety
- ▶ Design for the Value of Sustainability
- ▶ Design for Values and the Definition, Specification, and Operationalization of Values
- ▶ Design for Values in Healthcare Technology
- ▶ Design for Values in ICT
- ▶ Design Methods in Design for Values
- ▶ Modeling for Design for Values
- ▶ Participatory Design and Design for Values
- ▶ Technology Assessment and Design for Values
- ▶ Value Sensitive Design: Applications, Adaptations, and Critiques

## References

Abbott A (1988) The system of professions. An essay on the division of expert labor. University of Chicago Press, Chicago

Akao Y (ed) (1990) Quality function deployment. Integrating customer requirements into product design. Productivity Press, Cambridge, MA

Albrechtslund A (2007) Ethics and technology design. Ethics Inf Technol 9:63–72

Alexander JK (2009) The concept of efficiency: an historical analysis. In: Meijers A (ed) Handbook of the philosophy of science, vol 9, Philosophy of technology and engineering sciences. Elsevier, Oxford, pp 1007–1030

Asveld L, Roeser S (eds) (2009) The ethics of technological risk. Earthscan, London

Bhamra T, Lofthouse V (2007) Design for sustainability: a practical approach. Gower, Aldershot

Birkeland J (2002) Design for sustainability. A source book for ecological integrated solutions. Earthscan, London

Bogue R (2007) Design for disassembly: a critical twenty-first century discipline. Assembly Autom 27(4):285–289. doi:10.1108/01445150710827069

Bogue R (2012) Design for manufacture and assembly: background, capabilities and applications. Assembly Autom 32(2):112–118. doi:10.1108/01445151211212262

Boothroyd G, Dewhurst P, Knight WA (2011) Product design for manufacture and assembly, 3rd edn, Manufacturing engineering and materials processing. CRC Press, Boca Raton

Bucciarelli LL (1994) Designing engineers. MIT Press, Cambridge, MA

Calhoun DH (1960) The American civil engineer. Origins and conflict. MIT Press, Cambridge, MA

Calvert MA (1967) The mechanical engineer in America. 1830–1910. John Hopkins University Press, Baltimore

Chang R (ed) (1997) Incommensurability, incomparability, and practical reasoning. Harvard University Press, Cambridge, MA

Cheng T (2009) Design for reliability and robustness. IEEE Des Test Comput 26(6):2–3

Clarkson J (2003) Inclusive design: design for the whole population. Springer, London/New York

Cross N (2008) Engineering design methods. Strategies for product design, 4th edn. Wiley, Chichester

Cummings ML (2006) Integrating ethics in design through the value-sensitive design approach. Sci Eng Ethics 12:701–715

Davis M (1998) Thinking like an engineer. Studies in the ethics of a profession. Oxford University Press, New York/Oxford

De Vries MJ (2009) Translating customer requirements into technical specifications. In: Meijers A (ed) Handbook of the philosophy of science, vol 9, Philosophy of technology and engineering sciences. Elsevier, Oxford, pp 489–512

Desmet PMA, Pohlmeyer AE (2013) Positive design: an introduction to design for subjective well-being. Int J Des 7(3):5–19

Dym CL, Little P, Orwin EJ (2014) Engineering design: a project-based introduction, 4th edn. Wiley, New York

Eekels J, Roozenburg NFM (1991) A methodological comparison of the structures of scientific research and engineering design. Their similarities and differences. Des Stud 12(4):197–203

Erlandson RF (2008) Universal and accessible design for products, services, and processes. CRC Press, Boca Raton

Feenberg A (1995) Alternative modernity: the technical turn in philosophy and social theory. University of California Press, Berkeley

Feng P (2000) Rethinking technology, revitalizing ethics: overcoming barriers to ethical design. Sci Eng Ethics 6(2):207–220

Flanagan M, Howe DC, Nissenbaum H (2008) Embodying values in technology. Theory and practise. In: Van den Hoven J, Weckert J (eds) Information technology and moral philosophy. Cambridge University Press, Cambridge, pp 322–353

Frankena WK (1973) Ethics, 2nd edn. Prentice Hall, Englewood Cliffs
Franssen M (2005) Arrow's theorem, multi-criteria decision problems and multi-attribute preferences in engineering design. Res Eng Des 16:42–56
Friedman B (1996) Value-sensitive design. Interactions 3(6):17–23
Friedman B, Kahn PHJ (2003) Human values, ethics and design. In: Jacko J, Sears A (eds) Handbook of human-computer interaction. Lawrence Erlbaum Associates, Mahwah, pp 1177–1201
Friedman B, Kahn PHJ, Borning A (2006) Value sensitive design and information systems. In: Zhang P, Galletta D (eds) Human-computer interaction in management information systems: foundations, vol 5, Advances in management information systems. M.E Sharpe, Armonk, pp 348–372
Gaustad G, Olivetti E, Kirchain R (2010) Design for recycling. J Ind Ecol 14(2):286–308. doi:10.1111/j.1530-9290.2010.00229.x
Govers CPM (1996) What and how about quality function deployment (QFD). International Journal of Production Economics 46–47:575–585
Grady JO (1993) Systems requirements analysis. McGraw-Hill, New York
Grin J, Hoppe R (1995) Toward a comparative framework for learning from experiences with interactive technology assessment. Org Environ 9(1):99–120. doi:10.1177/108602669500900105
Grin J, van der Graaf H (1996) Technology assessment as learning. Sci Technol Hum Values 21 (1):72–99
Grunwald A (2009) Technology assessment: concepts and methods. In: Meijers A (ed) Handbook of the philosophy of science, vol 9, Philosophy of technology and engineering sciences. Elsevier, Oxford, pp 1103–1146
Guston DH, Sarewitz D (2002) Real-time technology assessment. Technol Soc 24(1-2):93–109. doi:10.1016/s0160-791x(01)00047-1
Hansson SO (2003) Ethical criteria of risk acceptance. Erkenntnis 59:291–309
Hansson SO (2007) Safe design. Techne 10(1):43–49
Hansson SO (2009) Risk and safety in technology. In: Meijers A (ed) Handbook of the philosophy of science, vol 9, Philosophy of technology and engineering sciences. Elsevier, Oxford, pp 1069–1102
Harris CE, Pritchard MS, Rabins MJ (2013) Engineering ethics: concepts and cases, 5th edn. Wadsworth – Cengage, Boston
Hauser JR, Clausing D (1988) The house of quality. Harv Bus Rev 66(3):63–73
Henderson K (1991) Flexible sketches and inflexible data bases: visual communication, conscription devices, and boundary objects in design engineering. Sci Technol Hum Values 16 (4):448–473
Holt R, Barnes C (2010) Towards an integrated approach to "Design for X": an agenda for decision-based DFX research. Res Eng Des 21(2):123–136
Hubka V (1982) Principles of engineering design. Butterworth Scientific, London
Hull E, Jackson K, Dick J (2005) Requirements engineering, 2nd edn. Springer, London
Imrie R, Hall P (2001) Inclusive design : designing and developing accessible environments. Spon Press, London;New York
Ireson WG, Coombs CF, Moss RY (1996) Handbook of reliability engineering and management, 2nd edn. McGraw Hill, New York
Jordan PW (2003) Designing pleasurable products: an introduction to the new human factors. Taylor & Francis, London
Keates S, Clarkson J (2003) Countering design exclusion: an introduction to inclusive design. Springer, London
Keeney RL (1992) Value-focused thinking: a path to creative decisionmaking. Harvard University Press, Cambridge, MA
Keeney RL, Raiffa H (1993) Decisions with multiple objectives: preferences and value tradeoffs. Cambridge University Press, Cambridge, UK/New York

King B (1989) Better designs in half the time. Implementing quality function deployment in America, 3rd edn. Goal/QPC, Methuen, MA
Koskinen I, Battarbee K, Mattelmäki T (eds) (2003) Emphatic design. User experience in product design. IT Press, Helsinki
Kroes P (2010) Engineering and the dual nature of technical artefacts. Camb J Econ 34(1):51–62. doi:10.1093/cje/bep019
Kroes P, Verbeek P-P (eds) (2014) The moral status of technical artefacts. Springer, Dordrecht
Kuo T-C, Huang SH, Zhang H-C (2001) Design for manufacture and design for 'X': concepts, applications, and perspectives. Comput Ind Eng 41(3):241–260. doi:10.1016/s0360-8352(01)00045-6
Lintsen H (1985) Ingenieur van Beroep. Historie, Praktijk, Macht en Opvattingen van Ingenieurs in Nederland. Ingenieurspers, Den Haag
Manders-Huits N (2011) What values in design? The challenge of incorporating moral values into design. Sci Eng Ethics 17(2):271–287. doi:10.1007/s11948-010-9198-2
Manohar K, Ishii K (2009) Design for supply chain: evaluation of supply chain metrics. In: IMECE2008: proceedings of the ASME international mechanical engineering congress and exposition, Boston, Massachusetts, USA, vol 4, pp 203–211
Nieusma D (2004) Alternative design scholarship: working toward appropriate design. Des Issues 20(3):13–24
Oosterlaken I (2009) Design for development: a capability approach. Des Issues 25(4):91–102
Pahl G, Beitz W, Feldhusen J, Grote K-H (2007) Engineering design: a systematic approach (trans: KWaLBTa), 3rd edn. Springer, London
Papanek VJ (1984) Design for the real world: human ecology and social change, 2nd edn. Van Nostrand Reinhold, New York
Papanek VJ (1995) The green imperative: natural design for the real world. Thames and Hudson, New York
Park SH, Antony J (2008) Robust design for quality engineering and Six Sigma. World Scientific, Singapore/Hackensack
Preiser WFE, Ostroff E (2001) Universal design handbook. McGraw-Hill, New York
Pritchard MS (2009) Professional standards in engineering practice. In: Meijers A (ed) Handbook of the philosophy of science, vol 9, Philosophy of technology and engineering sciences. Elsevier, Oxford, pp 953–971
Pugh S (1991) Total design: integrated methods for successful product engineering. Addison-Wesley, Wokingham/Reading
Raheja D, Gullo LJ (2012) Design for reliability, Wiley series in quality & reliability engineering. Wiley/IEEE Press, Hoboken
Raz J (1986) The morality of freedom. Oxford University Press, Oxford
Reuzel RPB, van der Wilt GJ, ten Have HAML, de Vries Robb PF (2001) Interactive technology assessment and wide reflective equilibrium. J Med Philos Forum Bioeth Philos Med 26(3):245–261
Reynolds TS (1983) 75 years of progress. A history of the American Institute of Chemical Engineers, 1908–1983. The Institute, New York
Rip A, Misa T, Schot J (eds) (1995) Managing technology in society. The approach of constructive technology assessment. Pinter, London/New York
Roozenburg N, Cross N (1991) Models of the design process—integrating across the disciplines. In: International Conference on Engineering Design (ICED-91), Zurich, pp 186–193.
Schot J, Rip A (1997) The past and future of constructive technology assessment. Technol Forecast Soc Change 54(2/3):251–268
Sclove RE (1995) Democracy and technology. The Guilford Press, New York
Shrader-Frechette KS (1991) Risk and rationality. Philosophical foundations for populist reform. University of California Press, Berkeley
Slater JA (2000) Design for maintenance. In: Quality Reliability and Maintenance (Qrm 2000), pp 253–256

Stauffer LA, Ullman DG (1988) A comparison of the results of empirical studies into the mechanical design process. Des Stud 9(2):107–114
Stauffer LA, Ullman DG, Dietterich TG (1987) Protocol analysis of mechanical engineering design. In: Eder WE (ed) Proceedings of the 1987 international conference on engineering design, ICED 87, Boston
Tatum JS (2004) The challenge of responsible design. Des Issues 20(3):66–80. doi:10.1162/0747936041423307
Ullman DG (1997) The mechanical design process. McGraw-Hill, New York
Ullman DG, Stauffer LA, Dietterich TG (1987) Toward Expert CAD. Comput Mech Eng November/December 1987:56–70
Van de Poel I (2007) Methodological problems in QFD and directions for future development. Res Eng Des 18(1):21–36
Van de Poel I (2009) Values in engineering design. In: Meijers A (ed) Handbook of the philosophy of science, vol 9, Philosophy of technology and engineering sciences. Elsevier, Oxford, pp 973–1006
Van de Poel I (2010) Philosophy and engineering: setting the stage. In: Van de Poel I, Goldberg DE (eds) Philosophy and engineering. An emerging agenda. Springer, Dordrecht, pp 1–11
Van de Poel I (2012) Can we design for well-being? In: Brey P, Briggle A, Spence E (eds) The good life in a technological age. Routledge, New York, pp 295–306
Van de Poel I (2013) Translating values into design requirements. In: Mitchfelder D, McCarty N, Goldberg DE (eds) Philosophy and engineering: reflections on practice, principles and process. Springer, Dordrecht, pp 253–266
Van de Poel I, Royakkers L (2011) Ethics, technology and engineering. Wiley-Blackwell, Oxford
Van den Hoven J (2005) Design for values and values for design. Inf Age 7(2):4–7
Van der Ryn S, Cowan S (2007) Ecological design, 10th anniversary edn. Island Press, Washington, DC
Vermaas P, Kroes P, van de Poel I, Franssen M, Houkes W (2011) A Philosophy of technology: from technical artefacts to sociotechnical systems, vol 6. Synthesis lectures on engineers, technology and society, vol 1. doi:10.2200/S00321ED1V01Y201012ETS014
Visser W (1990) More or less following a plan during design: opportunistic deviations in specification. Int J Man-Machine Stud 33:247–278
Visser W (2009) Design: one, but in different forms. Des Stud 30:187–223
WCED (1987) Our common future. Report of the World Commission on Environment and Development. Oxford University Press, Oxford
Whiteley N (1993) Design for society. Reaktion Books, London
Winner L (1986) The whale and the reactor; a search for the limits in an age of high technology. The University of Chicago Press, Chicago/London
World Health Organization (2006) Constitution of the World Health Organization – basic documents, supplement, Forty-fifth edn
Young RR (2003) The requirements engineering handbook, Artech House technology management and professional development library. Artech House, Boston

# Design for Values in the Fashion and Textile Industry

Claudia Eckert

## Contents



**Abstract**

The fashion and textile industry is one of the largest industries in the world producing billions of garments every year with a remarkably low awareness of the moral issues associated with the production and use of garments. After a brief introduction to fashion as a cultural phenomenon, this chapter explains the life cycle of garment production and use, which uses large amounts of energy and water and deploys many toxic chemicals. Globalized production raises many

C. Eckert (✉)
Engineering and Innovation, The Open University, Milton Keyes, UK
e-mail: claudia.eckert@open.ac.uk



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_37

issues around the ethical employment of staff. Design decisions have to be taken throughout the life cycle, but are often highly constrained by the commercial pressures of an industry with very low profit margins. Making moral decision in design is therefore in many cases a selection of the least harmful option. However, the chapter explains how some designers have found business models that allow them to produce garments in a least harmful way. The chapter concludes with a brief discussion of the conflicting drivers in design for value in the fashion and textile industry.

**Keywords**

Fashion and textiles • Values • Sustainability • Ethical production • Ethical consumption

## Introduction

We all own scores of garments at any one time. We dispose of many clothes before they reach their natural end of life, because they go out of fashion, they do not fit any more, or we simply do not like them anymore. An individual garment does not have an enormous social or environmental impact, either in how it is produced or how it is used. But the cumulative effect of the billions of garments that are produced, used, and disposed of worldwide is enormous, on individual lives, on national economies, and on the environment in large areas of the world. This raises the question of what values are expressed in the production of all these clothes, could they be different, and what effect might designing clothes for different values have on the world?

Designing, manufacturing, and selling clothes and other textile products involve the expression of two distinct types of values: the values expressed through the clothes and the values concerning the creation of clothes. The more conspicuous of these are the values that the purchasers and wearers of clothes express through their style choices and are enabled or prevented from expressing by the range of clothes offered to them. The role of designers and design choices in the construction and expression of styles and the values implicit in them is beyond the remit of this chapter. However the values expressed in the style choices of consumers influence and are influenced by *how* garments and other textile products are designed, manufactured, and sold. This chapter concentrates on designing for implicit or explicit values in the life cycle of making, using, and disposing of clothes in the fashion and textile industry.

The production of clothes and other textile products involves trade-offs between factors that may be viewed as values. How far it is possible for individuals and organizations to make value-driven choices, and how far their actions are dictated by economic and political necessity, is a complicated question. The main countervailing force is of course cost and the need to make a profit in a keenly competitive market, and some positive values may be in conflict with this force. These value conflicts (real or hypothetical) include:

- Improving quality and durability of the product itself; some design choices will increase the robustness and durability of the product without this being visible as added value to most consumers.
- Producing products that can be recycled easily or cheaply.
- Producing in one location rather than another, for instance, to employ staff in the country in which the company is based or to move parts of the production to low-wage countries.
- Maintaining long-term relationships with supplier companies.
- Paying above the minimum possible rates of pay.
- Ensuring better working conditions for the employees of supplier companies, or insisting on health and safety standards comparable to those required by law in Europe or North America.
- Minimizing the use of energy, water, or other scarce resources in production.
- Minimizing the pollution resulting from production.
- Minimizing the environmental degradation resulting from the production of raw materials, notably by making use of organic cotton.

Producing clothes and other textile products involves a variety of design activities, not all of which are seen as design within the fashion industry. There are two distinct groups of designers who influence the products and the effects they have: the fashion or textile designers and the designers of enabling technologies such as knitting machines and printers.

Fashion and textile designers specify the visual and tactile appearance of yarns, fabrics, and the products made from them. However, they typically operate within tight constraints imposed by cost-driven manufacturing and profit-driven retailers. The fashion and textile industry itself has a very narrow definition of design, only acknowledging the definition of visual and tactile properties as design. This excludes, for example, the knitting machine technicians who carry out the detailed design of knitted garments, which determines the details of the production process and thus aspects like the energy consumption in production or the waste of fabric (see Eckert 2001; Eckert and Stacey 2014).

Design for values in the fashion industry is often about choosing the less harmful option from a set of aesthetically equivalent alternatives and influencing manufacturers, retailers, and consumers to adopt more ethical alternatives. Very few designers are empowered to create products that are as sustainable and as ethical as possible. There is a small but visible group of fashion designers who create collections in the most sustainable and ethical way possible, distancing themselves from the bulk of the textile industry. These designers serve as role models for others and have advanced the debate on values in fashion, but have made very little impact on the environmental and ethical impact of the fashion industry at large. As their products are considerably more expensive than most comparable garments, they typically cater for a market of ethically conscious affluent customers, who use these eco-brands to express their own values and group membership. Much ethical and sustainable fashion is produced by micro companies that like other designer-led fashion business usually earn far less money than their public profile would suggest;

scaling sustainable and ethical design businesses is proving a great challenge in a globalized world.

Another group of designers that have a huge impact on environmental sustainability and ethical production are designers of the enabling technologies for garment production. The fundamental principles behind most textile production technologies have changed very little. However modern machines have become extremely reliable and versatile; and current innovations focus on optimizing energy consumption as a means to gain competitive advantage. Similarly the chemical engineers working on dye stuffs and finishes have been working on reducing the impact of the chemicals they are using. It would also be possible to think of the design of supply chains in textile production as a design problem, with the logistics experts as designers. The key choices affecting values and profits are almost always management decisions rather than part of design as it is usually understood.

Fashion is a system of continuous renewal and obsolescence of objects that are functionally equivalent and become desirable because they are new, different, or exclusive. Trends in fashion respond to cultural phenomena and often follow the leads of celebrities or star designers. However, mainstream high street fashion is a highly self-referential system where designers pick up on trends they see in the garments coming on sale and incorporate them into their own styles to create products that look both different to attract the customers and sufficiently similar to fit in with the other clothes on sale and in people's wardrobes. As these trends emerge, some are more or less sustainable or ethical; for example, when hand embroidery is part of a trend, a large number of people have to be hired who are no longer needed when fashion has moved on. While designers create the detailed appearance of their garments, they typically have to work within these trends.

While fashion and clothing is endlessly discussed in the media, fashion professionals rarely discuss their own processes or reflect on their practice in a formal and organized way. The discussion of moral issues, such as environmental sustainability or labor conditions in the developing world, is done either by academics or by campaigning groups who are outside the fashion industry, such as the fair-trade movement.

Companies appear to become active when they fear an adverse economic effect on their businesses. An example of this is Primark which has compensated victims and their families affected by the Rana Plaza factory collapse in Bangladesh[1] in 2013, where over 1,100 people died and 2,500 survived, 600 with severe injuries (Siegle and Burke 2014). The building had been designed as a lower building for shops and office to which additional floors were added, where factories with heavy equipment were housed. The building had shown cracks on the day before the collapse, but the textile factories forced their worked to work regardless of the apparent risk.[2] As this example has been widely discussed in the media, we will use it in this chapter.

[1] http://www.bbc.co.uk/news/business-24646942

[2] http://www.telegraph.co.uk/news/worldnews/asia/bangladesh/10036546/Bangladesh-Rana-Plaza-architect-says-building-was-never-meant-for-factories.html

The moral values associated with fashion and textile consumption are to some extent in direct conflict with each other, between individuals' needs to express themselves in ever-changing styles and the wastefulness of disposing of garments that could be worn for many more years, or between the desire to reduce the environmental impact of products by producing locally under controlled conditions in western countries and the desire of third-world countries to improve their collective economic situation through the employment of their workers in conditions that people in western countries would regard as exploitation through low pay, lack of legal protection, and suppression of union activities.

The chapter starts with a short historic and theoretical discussion of fashion and clothing in section "Fashion and Clothing" before providing a systemic view of the fashion life cycle in section "A Systemic View of Fashion". Section "Design for Values" discusses the two groups of designers who influence the fashion life cycle: the designers of the garments and the designers of the enabling technology. Section "Moral Responsibility in the Fashion and Textile Industry" posed the question who holds the moral responsibility for values in the fashion industry before section "Approaches to Ethical and Sustainable Fashion" provides examples of approaching these values before a discussion in section "Discussion" and conclusions in section "Conclusions".

## Fashion and Clothing

Clothing meets one of the fundamental human needs protecting us against the elements. Up to the industrial revolution in the nineteenth century, the majority of clothing has been produced locally. However textiles have been traded over very large distance since antiquity; for example, the Romans imported silk from China and cotton from India to supplement the wool and linen produced in the empire. The trade of raw materials and finished textile products has been a significant driving force of political events and wealth ever since and is part of a long and ignoble history of exploitation. Fibers and textile products are still traded globally. However a very significant change has occurred in how clothing is valued. Textile products used to be expensive compared to other commodities. Up to well into the twentieth century, people used to own a small number of clothes that they looked after well. Clothes were kept over long periods of time. The clothes were adapted to changing uses and styles or were mended when necessary. Clothes that became surplus to requirements were sold on or passed to poorer people. The fabric of worn out clothes was reused or recycled for paper. Now clothing has very much become a throwaway commodity: people have been buying an increasing number of clothing items which they only wear a few times before disposing of them. UK consumers spent £780 purchasing 35 kg of textiles in 2004 (Allwood et al. 2006).

The clothing, footwear, and textile sector is the fifth largest economic sector, employing up to 40 million worldwide, of which up to 19 million are employed in China, 2.7 million in the EU, and 400,000 in the UK (excluding retail), where it employs as many as the aerospace and automotive sectors combined (OECD 2004).

Between 2005 and 2011, the value of global apparel exports rose by 48 to USD 412 billion dollars. The top ten developing country suppliers now account for 58 % of global apparel exports, with China taking 37 % of that share in 2011 (IDE-JETRO 2013). Allwood et al. (2006) point out that in the UK, the amount of clothes purchased per capita grew between 2001 and 2005 by 37 %. As the price of clothing has fallen, the value end of the market (i.e., cheaper clothing) is booming, doubling in size in just five years to £6 billion of sales in 2005 (Lee 2007, p. 24).

Fashion is a phenomenon not limited to clothes and textile products, but is often associated most closely with clothing. Sproles and Burns (1994) define fashion as "a style of consumer product or way of behaving that is temporarily adopted by a discernible proportion of members of a social group because that chosen style or behaviour is perceived to be socially appropriate for the time and situation," whereas Welters and Lillethun (2011) define fashion, "as changing styles of dress and appearance that are adopted by a group of people at any given time and place." Sproles and Burns (1994) point out that fashion artifacts can be viewed as "symbols possessing meaning beyond their tangible characteristics." According to Flügel (1930) clothes serve three purposes: providing protection against the element plus the contradictory desires to display our bodies while covering them up, which he phrased as "decoration" and "modesty." Clothing can "communicate much more about the person than the social status he or she occupies or aspires to … gender, sexuality, age ascriptions, leisure inclinations, ethnic and religious identifications, political and ideological dispositions, and still other attributes of the person can be in play in the clothes we wear" (Davis 1992, p. 112). Fashion is a key means for people to express membership of a particular social group (Polhemus and Procter 1978, p. 20).

Simon-Miller (1985) makes the distinction between clothing as language, a "conventionalised set of norms that leads X to be perceived, by his clothes, as a businessman, or Y as a soldier," and fashion as speech, "a statement which leads Z to take Y's fatigues [i.e., military uniform] and accessorise them into a stylish, urban outfit." Buying new clothing signals both the ability to understand what is currently fashionable and the ability to afford clothing to keep up with changing trends. Over the last 50 years, the rate of change in fashion has accelerated greatly (Lowe and Lowe 1985); women's wear in particular sees multiple seasons per year for catwalk fashion, not to mention the growing and constantly evolving influence of street and subcultural styles. Even since 1985 much has changed: today's commercial fashion system is a highly developed complex of brands, designers, retailers, imitation, and adoption hierarchies, with the Internet accelerating information flows and feedback loops (Cappetta et al. 2006).

As Fig. 1 illustrates, products stay in fashion for a certain period of time, with new styles being added as others go out of fashion. Some designs outside of the space of fashion at any one time and either might not sell or won't be offered. Some of these designs might later come into fashion. Fashion designs are typically based on garments offered by stylistic leaders, which are than adapted and reinterpreted for a specific market (Eckert and Stacey 2000) rather than the needs and desires of

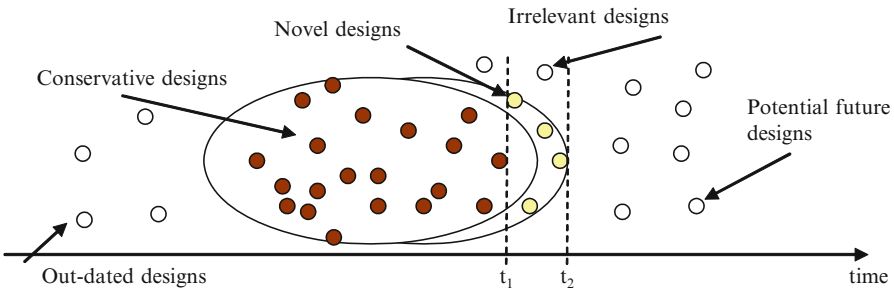



**Fig. 1** The envelop of accepted designs (From Eckert and Stacey 2001)

potential customers as would be the case in other design fields. Fashion tends to evolve with new features coming into fashion while others slowly fade out. Others are not yet acceptable to the market and might be picked up in future collections. Customers have very little input into design process and only influence the development of fashion through the styles that they choose from the garments on sale.

The changes in acceptable or preferred styles mean that many garments are discarded before the garments have reached their natural end of life. Fashion is a powerful economic driver, sustaining global industry and employment, but fashion's inbuilt obsolescence is intrinsically unsustainable – a contradiction at the heart of contemporary fashion consumption which Black has termed "The Fashion Paradox" (Black 2008). She acknowledges the need for fashion and with it the fast renewal of styles and advocates addressing sustainability in the production and distribution of garments while empowering the consumer to acquire garments that meet their personal needs in terms of style and in particular fit, as many garments are rarely worn because they no longer fit or never fitted. The Considerate Design project (Black and Eckert 2009) therefore explored mass customization as a means to improve sustainability. Fletcher (2008) questions the whole fashion system more fundamentally and advocates changes in both consumption and production, for example, by repatriating production to avoid the negative impact of transport.

## A Systemic View of Fashion

To understand the ethical and sustainability issues as associated with clothing, it is important to look at fashion and textiles as a global system operating across the entire life cycle of the garment. Many of the issues that concern clothing are shared with other consumer products, but aggravated in the case of fashion due to the sheer volume of clothes produced and consumed.

Until the 1980s most garments were designed and manufactured either in the same factory or fairly close together. Since then production has been moved repeatedly to cheaper production locations leading to lower production prices but also to reduce flexibility in the supply chain. For many countries textile production has been a stepping stone to later attracting higher-value production into the

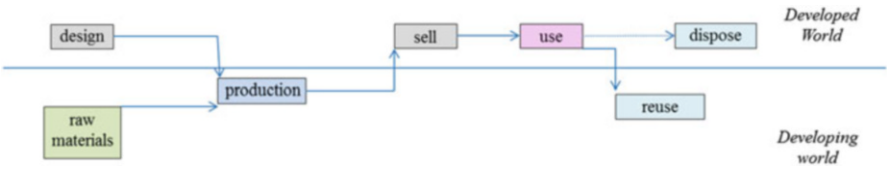



**Fig. 2** Life cycle of garments

country when textile production moves on to other places. However when the production was moved to cheaper location by customers, this has posed a significant problem for workers who lost their jobs and has to reskill. This was the case first for Eastern Europe and later for China. Currently places like Bangladesh and Vietnam, which have a workforce willing to work for very little money to escape from the poverty of an agricultural society, have become centers of textile production.

While some garments are produced and used in the same part of the world, most garments sold in the developed world follow the pattern shown in Fig. 2, of design where the majority of customers live and trends emerge, and production in cheap labor countries. The fashion mass market is dominated by large retail chains that have been competing on price. Asia is becoming an increasingly large market for western brands as well as new emerging Asian brands which also design in Asia. The European Union alone imported clothing products and merchandises worth $170,058.1 million (WTO 2013).

This section provides a systemic view of the life cycle of garments including alternatives for the major groups of products, first from the raw materials to the garment in the shop and then from the point of purchase to the end of life. Some of the ethical issues will be highlighted.

## The Production of Clothing

The supply chain of garments overlaps with supply chains of many other products and requires input from other industry sectors; it therefore shares the problems associated with them.

Figure 3 provides an overview of the major stages in the production of textiles and an indicative profile of the major environmental impacts of the different stages of the production life cycle. The production of textiles begins with the fibers from which the fabric is made. Fibers come from three main sources:

- Animal fibers, mainly wool from sheep but also cashmere, mohair, and angora.
- Vegetable fibers, mainly cotton and linen as well as fibers made from cellulose.
- Man-made fibers which are produced from fossil fuel.

Natural fibers are then cleaned and prepared before they are spun into threads for weaving and yarns for knitting. Knitwear is usually knitted directly into panels that are assembled into garments. For tailored clothing the threads are woven or

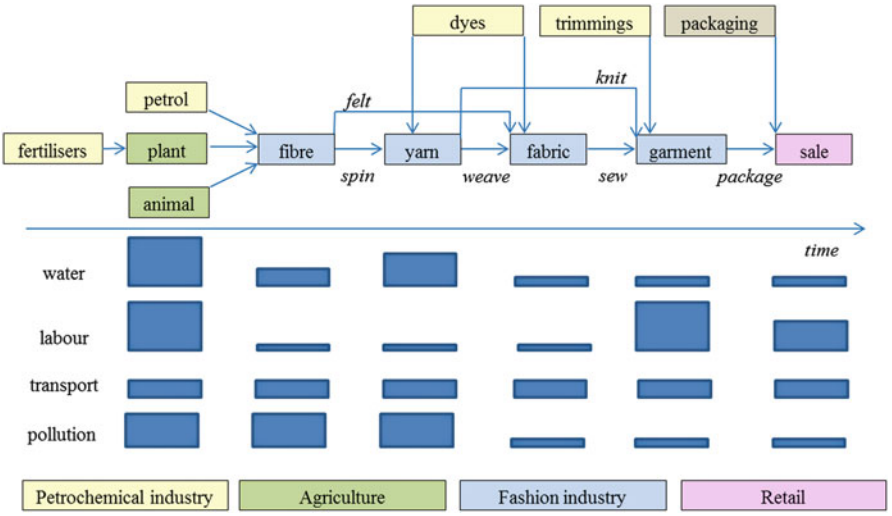



**Fig. 3** The production of textiles

knitted into fabrics, from which panels are cut and assembled into garments, which results in inevitable wastage of fabric. For so-called nonwoven fabrics, which are used for household fabrics like cleaning cloths, the fibers are directly formed into fabric.

The materials can be dyed at each stage, as a fiber, as fabric, or occasionally as garments. The newest development is digital printing straight onto fabric panels.

In the assembly process trimming, buttons, and zips are added. Many garments or fabrics are treated with chemicals after treatments to create a certain visual effect or to improve comfort or durability during wear. For example, many school uniform clothes or office clothes are now Teflon coated so that strains can be mopped up easily. The finished garments are then packaged and shipped to the shops for retail.

The transport of garments across the world can be considerable. For example, a designer interviewed as part of the Considerate Design project (Black and Eckert 2009) commented that she had worked for an Australian knitwear company which used Australian wool. The wool was shipped to Italy for spinning and then to China for knitting before being shipped back to Australia for packaging. The garments are sold as Australian wool garments worldwide. Textiles are mainly shipped in large container ships, which often travel slowly to save on fuel consumption. Garments are sold by fashion brands either directly or through retailers, who have their own distribution networks. The textile supply chain has significant markups between the different players in the supply chain. Allwood et al. (2006) provide the example of a classic white T-shirt where the cotton yarn from the USA costs £0.55, the knitted fabric from China £1.08, and the knitted T-shirt from China £1.96, with a UK wholesale price of £2.65 and a retail price of £7, showing that the largest share of the profit goes to the retailer.

The long delivery time means that the designs have to be finalized early and so might miss the taste of the market if predictions of the development of fashion are wrong. Up to the 1980s many garments were still produced fairly locally in Europe and transported by truck or rail. At that time it was possible to both reorder successful designs and cancel unsuccessful ones. Now garments are usually produced in preset numbers regardless of whether the design is successful and would sell in higher numbers or is unsuccessful and will end up in the sales or even as landfill.

## The Use and Disposal of Garments

In spite of endless coverage of fashion in fashion magazines and the mainstream press, surprisingly little is known about how textiles are consumed by the majority of the consumers. Figure 4 shows the remaining steps of the life cycle. Many garments are bought without ever being worn or being worn once. Besides the obvious examples like wedding dresses or ball gowns which are used for very limited occasions, many garments are not worn because they do not quite fit or the owner decides that they do not particularly like the garment after all. As clothes have become cheaper, anecdotal evidence also indicates that some – young woman in particular – buy clothes as almost deposable products to be worn once for an evening or event and then discarded and kept unused. WRAP (2012) point out that around 30 % of the clothing in the average UK household has not been worn for over a year. If garments are worn frequently, the majority of

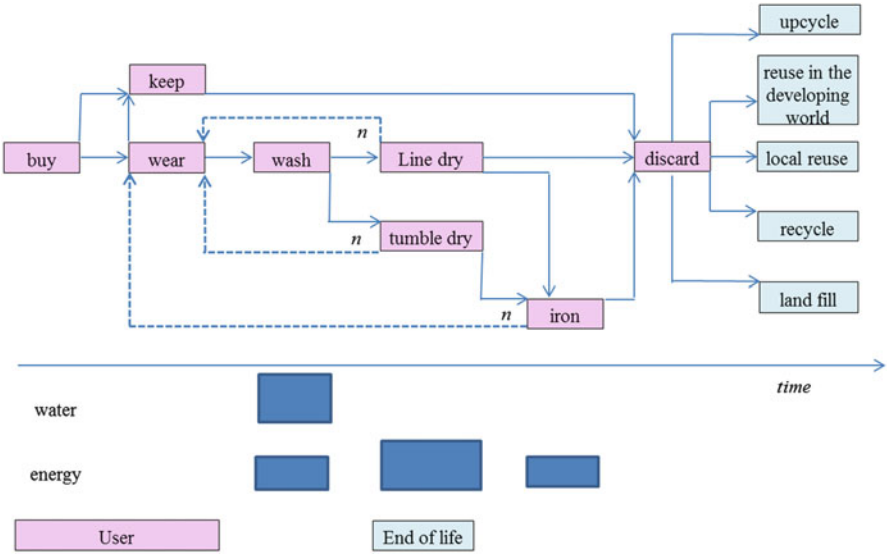



**Fig. 4** Use and end of life of garments

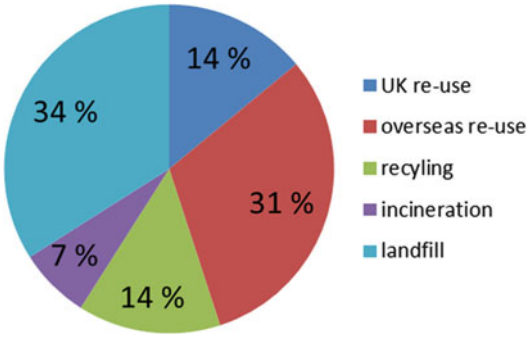



**Fig. 5** End of life of textile products in the UK, according to WRAP (2012)

their environmental impact lies in the use phase; otherwise the balance shifts to production.

If the garments are used, they need to be washed, dried, and maybe ironed. The energy and water use in these activities varies enormously with household practices and garment types. Some people wash garments in full 10 kg washing machines at 30 °C, while others wash their laundry in smaller loads and higher temperatures. Both tumble drying and ironing are to a large extent optional activities, which take up huge amounts of energy. When a garment needs to be washed is to some extent a matter of personal choice. People are concerned about smell, stains, and to overall appearance. While some people would wear the same garments for several days, others would change several times a day. Historically people washed clothes much less than we wash clothes now and used other garments like aprons to protect their main clothing.

Again there is a huge variation in how people discard garments (see Fig. 5). Some clothes are worn until they fall apart, while others are discarded in next to new condition. Some garments end up in landfill. Some local authorities collect garments for recycling. Garments can be recycled provided they are made of single materials. In practice this is rarely the case, as this would require removing fasteners such as buttons or zips. Often the threads or labels are made of different materials, thus limiting the possibility to recycle them. To avoid this label information is sometimes printed directly onto the inside of garments. Some garments are reused by passing them on to another person either within a family or friendship group, through charities, or secondhand trading such as eBay. Two thirds of UK consumers buy or receive secondhand clothing (WRAP 2012). Many UK charities, most famously Oxfam, sell secondhand garments to consumers in the same country as the donors both to raise money for the charities' general causes and to provide cheap clothing. Many used garments are shipped to the developing world.

While the steps described are generic, very little is known about the variation in behavior between individuals. There is no reliable data about how long people keep garments and how frequently they are washed. This makes it hard to estimate the number of garments owned and worn by people and therefore to deduce the environmental impact of garments. However a simple calculation can illustrate

the scale of the problem. If we make the (conservative) assumption that each person owns 10 T-shirts, then the whole of Britain owns at least 600 million T-shirts.

## Ethical Issues Associated with the Production of Clothing

All of the different stages of production of clothing have issues associated with them that they share with other goods that use similar processes. The main categories of problems that are usually discussed are the use of water and electricity as well as labor and transport issues.

Man-made fibers are produced from fossil fuels and therefore compete with other uses of plastic. Textile fibers can be produced from recycled materials, but are themselves difficult to recycle, because the material has to be uncontaminated, e.g., not mixed with wool and cotton. The issues here are the same that affect other recycling loops, which might cost more than they save. For example, fleece fabrics are can be produced from recycled plastic pallets. Germany has a sophisticated system of recycling of plastic bottles, which are brought to retailers, collected in collection centers, and turned into pallets. These pallets are bought up by Chinese manufacturers to make fleece fabric, because they are of high quality and comparatively cheap. They are shipped to China and often reexported back to Europe.

Arguably the biggest impact of textile production comes from the production of cotton, which requires enormous quantities of both water and fertilizer. The most extreme example is the Aral Sea, where years of diverting water for cotton production have resulted in the Aral Sea shrinking by over 80 % of its volume causing an environmental disaster (WWF 2008, p. 22). Cotton is competing with food crops, and the production of cotton for export is contributing to food shortages in Africa, in particular as the fertilizers required to grow cotton make the land unsuitable for arable use. Historically growing sheep for wool has transformed the landscapes as native landscapes have been transformed into grassland for grazing. This is particularly an issue in Australia and New Zealand, where the sheep are also using up precious water resources.

Spinning, weaving, and knitting are highly mechanized processes with high energy consumption. In this phase the major environmental impact lies in pollution arising from dyeing, printing, and finishing fabrics. State-of-the-art expensive technology is often not used in unregulated countries in the developing world.

The process of making tailored garment has been fundamentally unchanged for centuries. The fabric is cut and assembled by people on sewing machines. While parts of the process have been automated, complete automatic assembly remains expensive and dedicated to special applications. Sewing machine operators require a certain amount of skill and training, but very little by way of formal academic training and therefore can be recruited and trained swiftly in different parts of the world. Knitting machines have been developed over the last decade to knit garments, such as underwear, sportswear, or jumpers, in one piece; however these processes are still error prone, so that the assembly of knitted garments is still a labor-intensive process.

Finished garments are either quality checked and packaged at the manufacturer or shipped to specific processing plants. Fabric and garment inspection can now be largely automated, but as machinery is very expensive, inspections are largely carried out by human operators. Fashion mass production is highly competitive and operates with very low profit margins. Manufacturing jobs are monotonous and notoriously badly paid with operators being barely paid a living wage for long working hours and working under sometimes appalling conditions, as the recent problems in Bangladesh have shown.

The garments are then distributed to retailers across the world and therefore need to be packaged appropriately. Some are boxed as groups, while others are individually wrapped. A significant part of the impact of garments is in the retail: heating, lighting, and repairing the retail space eat resources. As these issues are shared with all other consumer goods, the problems and environmental impacts and the expression of values by various stakeholders involved in retailing are considered to be outside the scope of this chapter, but need to be considered.

In use the main sources of environmental impact are washing, drying, and ironing of clothes, which are greatly influenced by user behavior. Tumble drying and ironing are steps that can also be omitted entirely, thus changing the balance of impact from use to production. The disposal of garments shares many general issues associated with waste disposal. A particular issue in textiles is the fact that 31 % of textile waste is shipped to the developing world. While a small fraction is upcycled, i.e., turned into a higher-value product, most are sold at markets in the big cities, while the rest is sold by traveling salesmen in remote areas. This had a devastating effect on the indigenous textile industries, which cannot compete on price with the very cheap imported garments (Sinha et al. 2012).

In summary the ethical issues associated with the production of clothing can be associated with the following categories of problem, shown by the indicative profile in Fig. 3:

- Water being used in the course of the growing or creating the fibers and dyeing and finishing the fabric and garments and later in laundry.
- Pollution of drinking water and wastewater during production as well as the pollution through fertilizers for growing cotton and other vegetable crops.
- Energy used in the production of the garments and the transport of garments across the world as well as the energy used in washing, drying, and ironing clothes.
- Labor conditions for the workers involved in growing the raw materials and making the garments
- The effect the reexport of used garments has on production in the developing world.

The major occupational hazards associated with the textile industry are hazardous chemicals, fiber dust, noise, and monotonous repetitive processes (see Allwood et al. 2006).

## Design for Values

Many of the decisions that affect the ethical and environmental issues in the fashion and textile life cycle are commercial decisions taken for commercial reasons. Designers have to operate within these commercial constraints regardless where in the market they are operating and inevitably make compromises.

There are two categories of designers that have a huge impact of the fashion industry, the engineering designers who generate the enabling technology and the fashion and textile designers who design the individual garments. In the fashion industry the boundaries between commercial decisions and design decisions are blurred. In practice they are often associated with the organizational set of the companies more than the tasks that people carry out.

## Manufacturing Machine Builders

The fashion and textile industry has seen very few fundamental technological innovations over the last century. For example, the development of the first mechanical knitting machine dates back to William Lee's invention of the stocking frame in 1589, the spinning Jenny invented in 1769, and the Jacquard loom in 1801. The products they produce have also changed very little. Modern textile machines are highly optimized mature technology, which is adapted to particular uses. Production machinery has high energy consumption, and machine manufacturers have turned their attention to energy reduction as a distinguishing sales feature. For example, knitting needle manufacturers claim that they have been able to reduce energy consumption in the knitting process by 15 % by optimizing needle geometry, and knitting machine manufacturers claim reductions of 35 % up improving electronic control of the knitting process.

Technology is also used to merge previously separate production steps, for example, integrated spinning machines, which can move from raw fiber directly to threads or yarns, thus removing the need to transport materials between different operation steps, or seamless knitting machines which produce entire garments rather than individual panels to be sewn together or pieces of knitted fabric to be cut into shape.

Computer control technology has significantly reduced waste materials over the last two decades. Knitted garments are used to be cut from squares of fabric with significant offcuts, but are now mainly knitted into shaped panels, thus reducing the waste yarn. Laser printer technology has also reached the fashion industry. Both woven and knitted fabrics can now be printed to produce printed panels. Currently traditional printing methods, which use screens for individual colors, are still far more economical for large volumes. Laser printers have the potential to reduce waste fabric by enabling more flexible control over production volumes and reducing the amount of dye required.

The aim of the machine builders is to produce highly reliable machines to reduce the time when the machines are out of production and to reduce the number of

faulty garments, which become waste in production. Machine builders and the chemical industry are investing heavily in using less harmful chemicals.

## Fashion and Textile Designers

Designers are involved whenever there are aesthetic decisions involved (see Fig. 6). Yarn designers design yarns that are knitted and threads that are woven. Fabric designers (sometimes called textile designers) design weave and print patterns. Knitwear designers design knitwear and socks. Fashion designers design tailored garments, while contour fashion designers design underwear and swimwear. In the UK separate degree courses and specific training exist for each of these specialisms, but designers move between fields during their careers.

Practically there is a big difference between what design could do for values in the fashion industry and what designers can practically be able to do. With the exception of small designer-lead businesses, where the designer is involved in most decisions, designers work for brands or manufacturers with limited ability to make decisions about the business aspects including the price point. Designers work with business people on customer accounts and create products for specific customers and market segments, with known tastes and given price points. Suppliers are usually selected for commercial reasons and designers are mainly consulted on the quality of the product in terms of the suppliers' ability to meet design intent. Designers therefore have little direct influence on the ethical issues of production.

The designers' ability to make ethical and sustainable decisions is limited to the choices that they are able to make in their designs. The impact of garments can vary significantly between almost identical pieces depending on how and where it is produced. Designers know that and can select options that have less impact. For example, collections might include crochet clothes. This process cannot be done by a machine, and therefore garments have to be hand crocheted in the developing countries by workers who work on a piece commission bases. However, a similar visual effect can also be achieved by knitting a garment on a state-of-the-art

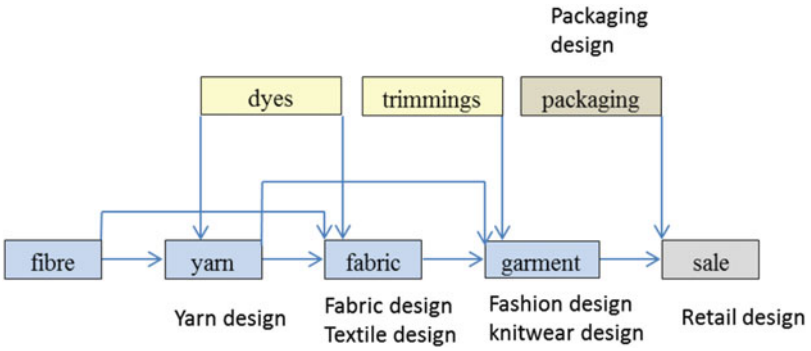



**Fig. 6** Design input in the fashion life cycle

machine. Designers also decide on the material composition of the fabrics or yarns that they are using; therefore they influence to some extent the durability of the garments and the way they need to be cared for. For example, putting an anticrease finish onto a shirt means that less energy will be used in ironing. Designers often have very little information concerning ethics and sustainability that they can use to select design options. At the moment there is a plethora of eco-labels, which are not standardized and only are applied by a fraction of producers (Sinha and Shah 2010). It is also very difficult to do a proper life cycle analysis of a garment, because the data is not available in standard life cycle assessment tools to cover the enormous variation in production and specialist analysis is not affordable considering the low profit margins of textiles.

Designers can also try to influence their organizations to adopt a more sustainable or ethical approach by pointing out either how this would enhance the brand or why this is necessary to compete in a particular market segment. Designers also influence in each through the products that they create. By starting or picking up on trends, they can create or encourage virtuous behavior. Many designers find their limited ability to influence their business and the repetitive and organizational nature of many design positions frustrating and move into more managerial roles, where they no longer generate the designs, but can work on market strategies, buyer management, or supplier selection.

## Moral Responsibility in the Fashion and Textile Industry

Legislation has a powerful impact on the production and use of garments by outlawing particular materials or practices, either by banning chemicals or processes used in production in its jurisdiction or by forbidding the sale of products under its jurisdiction that have been produced in a particular way. For example, European Union has banned the use of the chemical biocide dimethyl fumarate (DMF) for leather tanning in Europe in 1998 and banned the sale of products using the chemical in 2009.[3] However these bans can be difficult to enforce, since costly tests would be required to identify the chemicals in products. Therefore the onus of assuring the legislation is complied with falls to some extent to brands and retailers to assure that their supply chains are compliant.

Governments can also influence textile production by trade agreements. The World Trade Organization[4] introduced the Multifibre Agreement in 1974, which limited the import of textiles from developing countries into developed countries with the aim of enabling the developing countries to build up a textile industry for export. This was replaced in 1994 by a further agreement that brought textile incrementally in line with the general GATT rules applying to other products. The end of trade restriction has contributed to a significant fall in textile prices

[3] http://europa.eu/rapid/press-release_IP-09-190_en.htm

[4] http://www.wto.org/english/tratop_e/texti_e/texintro_e.htm

across the world over the last two decades and has been blamed for some of the increased consumption in the west (Allwood et al. 2006).

Governments of course also have the usual diplomatic and political means of influencing production and work conditions in other countries, as well as raising awareness in their own country. In 2011 the UK Department for Environment, Food and Rural Affairs commissioned a sustainable clothing road map in 2011 (Defra 2011), showing best practice and making suggestions of how to make clothing more sustainable through improving environmental performance, affecting consumer behavior, and creating public awareness.

The question of who is responsible for moral decisions in the fashion industry is hotly debated; and there is neither a simple answer nor a simple consensus. The issues around exploitation of workers have been debated in the public sphere in response to the collapse of the Rana Plaza building. The *Guardian* newspaper in the UK reported on the Rana Plaza collapse[5] and published a week later a summary table of the online discussion in response to this article. 36 % of the Guardian respondents blamed the corporations/retailers, 26 % the consumers, 22 % the factory owners/building inspectors, and 16 % the government.[6] While this is by no means a rigorous study or a statistically representative result, it highlights how divided public opinion is on the moral ownership of the problem. Are the consumers demanding too low prices or are the retailers pushing the suppliers to unreasonable price points? Are greedy manufacturers not paying adequate wages or making their workers work under unreasonable conditions? Are corrupt officials signing off unsafe building? The answer is probably yes to each of these questions. However this is not a deterministic system. Each of the players could to some extent push back and thereby cause a change in the system. Consumers have been paying higher prices for other goods, such as food. Workers in other industry sectors work across the world in far better working conditions under better pay.

In 2009 the Considerate Design project (see Black and Eckert 2009) looked at how garments might be designed so that they better meet the needs and desires of consumers while limiting their environmental impact. As part of the project, we conducted a workshop with over 20 industry and academia experts in London to identify the factors that affect this and identified a range of similar broad issues as shown in Fig. 7 (Eskandarypur et al. 2009). The workshop identified both barriers and enablers for sustainable and ethical design, which designers needed to become aware of in their practice. One of the main issues was raising awareness of values in fashion across the general public, who remain largely unaware of the issues and therefore not willing to accept any cost increases arising from responses to sustainability and ethical issues. Greater transparency of the impact through labeling and life cycle assessment tools was identified as a requirement not just for changing public attitudes but also enabling designers to make more sustainable decisions.

[5]http://www.theguardian.com/world/2014/apr/19/rana-plaza-bangladesh-one-year-on

[6]Guardian, Guardian weekend, 26.4.14, p. 8

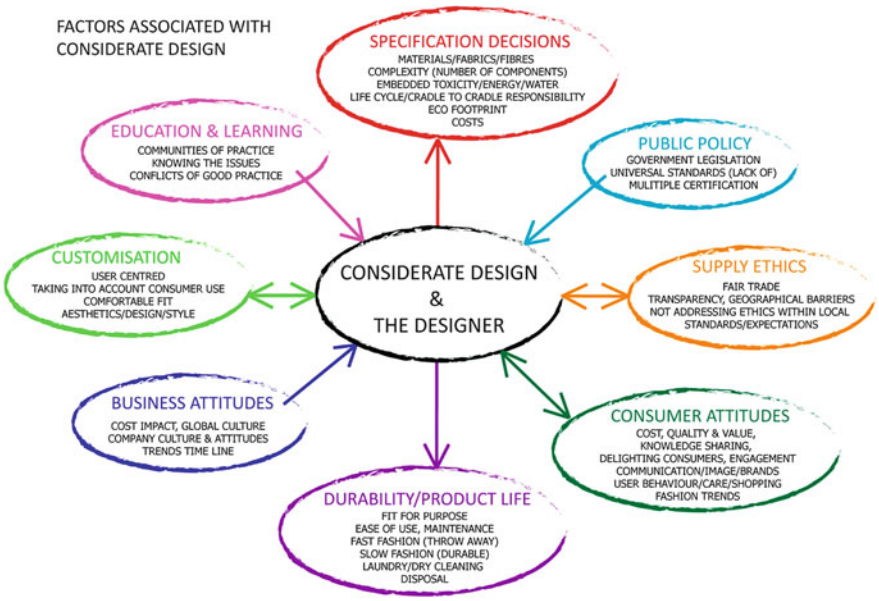



**Fig. 7** Systemic influences on considerate fashion

## Approaches to Ethical and Sustainable Fashion

Different stakeholders have taken their own approaches to creating more sustainable fashion products. The following section provides a flavor of the range of different responses to a common and systemic problem, which play out at different scales.

### The Ethical Product

Some designers and brands have decided to produce as sustainable fashion as possible by both sourcing and producing fashion as sustainably and as ethically as possible. This happens at very different scales.

Designers who produce their own products in many ways have the most control over their production. For example, Steven Harkin runs a small company for high-quality leatherwear.[7] He designs his own products and sources his leather from Italy from sustainable sources looking at both sustainable conditions for the animals and environmentally friendly tanning and dying processes. He makes the bags himself in the UK and sells them directly to retailers and the general public through trade

[7] http://www.stevenharkin.com/about-1-w.asp

fairs and makes the bags to order. He also takes the bags back and repairs them if necessary. While he has very little waste in his process, his bags are expensive using expensive high-quality materials. As he is running his own business with limited help, he often works overtime in particular to meet the deadlines of promised deliveries (see Wynn et al. 2011). While this is in many ways a very sustainable form of production, he is exploiting himself as a worker. Buying small orders he has to trust his suppliers and cannot personally control their processes. Steven Harkin's products look very smart and professional; other designers include a certain "eco-aesthetic" in the styling of their clothes. This has become a real problem as anecdotal evidence indicates that customers consciously avoid garments with an "eco"-look and feel. This appears to be a problem for Fairtrade® garments.

Sustainable production can also be a business model, as illustrated by the Californian sportswear company Patagonia, which has the company motto: "Build the best product, cause no unnecessary harm, use business to inspire and implement solutions to the environmental crisis."[8] The company uses materials from sustainable sources as much as possible, such as organic cotton, or recycled materials for man-made fibers. They monitor their suppliers' working conditions and publish their supply chain on their company Web pages. They aim to use energy from renewable sources as much as possible in their production and distribution. The garments are designed to be long lasting and repairable. At the end of life, the garments can be recycled through the stores. The price point of Patagonia clothing is at the upper end of high street clothing. Their products are aimed at the outdoor wear or sportswear market, where consumers are interested in performance and are typically willing to invest more and give more thought to purchasing products than for day to day clothing. Their customers being into nature and outdoor activities are to some extent naturally predisposed to care about environmental issues, and Patagonia has been very clever at using their environmental credentials in their marketing, so that they have been a brand of choice for a segment of the population.

Some companies produce products in a sustainable way, but do not market the fact because they do not want to put potential customers off with an eco-image for their products. For example, John Smedley[9] has been producing knitwear in Derbyshire in the UK since 1784. They specialize in fine gauge classical knitwear often used in catwalk shows. Almost all their designs are either made of New Zealand merino wool or Sea Island cotton, which is a stable long-fiber cotton, which is sourced "responsibly" from selected suppliers with whom they have long-standing relationships. They ship the raw materials in bulk to Derby and spin, dye, and finish the garments on site. This gives them not only much reduced transport but also full control over their production runs as they can respond in their production of yarn to market demands and stop unsuccessful production runs.

[8] http://www.patagonia.com/eu/enGB/patagonia.go?assetid=2047

[9] http://www.johnsmedley.com/uk/

Large clothing retailers are also increasingly embracing a sustainable agenda, as part of their corporate social responsibility as well as a means of generating a distinctive brand image. In 2007 Marks and Spencer, one of the largest retail chains for clothing in the UK, which for many years has provided a de facto benchmark for textile quality in the UK, launched the so-called Plan A across its entire product range,[10] which includes a combination of measures to improve ethical and sustainable issues across the fashion and textile production use system. This includes diverse activities like a project for the protection of the rain forest in Peru, health training for workers in Cambodia, and financial training for workers in India. In the UK Marks and Spencer have a clothing swapping scheme, where customers get M&S vouchers if they bring M&S garments to Oxfam charity shops. They also invested in aerodynamic lorries. The companies see this approach as a means to achieving both environmental and financial sustainability.

## Self-Regulation

Corporate social responsibility (CSR) means "companies integrate social and environmental concerns in their business operations and in their interactions with their stakeholders on a voluntary basis" by "going beyond the legal expectations and investing 'more' into human capital, the environment and the relations with stakeholders" (EC 2001). European policies have pushed CSR in different industry sectors through sectoral social dialogue committees. For textiles and clothing, they recommended core labor standards in 1997. As with all voluntary arrangement, the challenge lies in monitoring and enforcing policies. Nordestgaard and Kirton-Darling (2004) cite Patrick Itschert, the then General Secretary of the European Trade Union Federation for Textiles, Clothing and Leather ETUF-TCL, as saying that "beyond the ideal solution of having shop stewards in every plant monitoring agreed codes, it is possible to identify a number of application tools available to the trade union signatories of these codes of conduct: (a) involving international verification bodies (e.g., ILO, OECD etc.), (b) 'naming and shaming' companies who violate the codes signed, and (c) convincing market leaders to develop and participate in joint projects and apply peer pressure to other companies." These are exactly the mechanisms that have come into play after the Bangladeshi Rana Plaza building collapse. Western customers were identified and named. Some of the western companies have grouped together to set up compensation funds for the workers and their families.

Two legally binding agreements between western customers and Bangladeshi clothing manufacturers have been put into place. The Accord on Fire and Building Safety in Bangladesh[11] has been signed at date by 150 apparel corporations, two global trade unions, trade unions in Bangladesh, and several NGOs. The Alliance

[10] http://plana.marksandspencer.com/about

[11] http://www.bangladeshaccord.org/

for Bangladesh Workers Safety[12] is similar with mainly US companies involved. Bangladeshi companies also have these CSR strategies in place in response to the Rana Plaza disaster to assure their links to their customers; however these appear to be falling short of international expectations. Despite international concerns about workplace conditions and safety, their accountability and transparency still fall short of expectations according to Kamal and Deegan (2013). These agreements were initiated by the customers who were concerned by their reputation and wanted to reduce their burden of needing to inspect work places, rather than proactively generated by Bangladeshi companies or government.

### Behavioral Change

There are few in-depth studies of consumer behavior in fashion and textiles. Preliminary studies indicate a low awareness of sustainability and ethical issues in fashion in textiles (Crommentuijn-Marsh et al. 2010). In on-going studies participants with some awareness of fashion sustainability issues mentioned a few common sense approaches.

- Recycling: People both bought secondhand clothes, for example, on eBay, and recycled their own clothes through charity shops. Having purchased a product secondhand appeared to give the people the sense of being excused from the environmental impact of the garment.
- Eco-brands: They bought a limited amount of eco-brand and fair-trade garments and looked for organic cotton in high street brands. However, they also commented that since these garments are more expensive and offered limited choice, they only bought a fraction of their garments in this way.
- Conscious shopping: They tried to only buy garments if they felt they really needed them and select the garments carefully. Therefore they also selected classical styles.
- Made do and mend: They tried to look after garments and wear them until they are totally worn out.

However, it is difficult to say conclusively whether the strategies are driven by environmental and ethical concerns or by economic considerations.

## Discussion

The examples have shown that it is possible for individual stakeholders to operate in a way that is ethical and sustainable for them. However, there are fundamentally conflicting values at the heart of the fashion and textile industry, which make it difficult to make moral judgments.

[12] http://www.bangladeshworkersafety.org/

Changing fashion is a reflection of our changing culture, in which individuals express themselves consciously or unconsciously. We tire of clothes as they go out of fashion and like new and exciting things. Therefore many clothes are replaced before they are worn out and people own many more clothes than they would strictly need. The resulting level of production is on the long run not environmentally sustainable. The ability to afford new fashion items is an outward symbol of prosperity. We find high levels of consumption of cheap fashion among relatively poor people, who use clothing as a means to keep up. The same appears to apply to people in the developing world, which has already resulted in enormous growth in East Asian fashion markets and will aggravate the sustainability problem further on a global level. There is a direct conflict between sustainability in textiles and clothing and the right of self-expression.

Another major conflict of values lies between reduced and sustainable production in developed countries and the need of the developing world to build up an industry that provides employment for its growing population. By western standards the working conditions in textile factories in the developing world are extremely harsh with people working at low pay for long hours in very repetitive tasks. However, for the workers it is an opportunity to avoid even worse living conditions. The workers do not want us to move production back to western countries, but offer them better wages and working conditions. By the cruel logic of the economics of the textile industry, higher wages means the production will be moved to another place. China has now reached a point in its industrial development that it wants to move to higher-value jobs and lets low-wage job go, but other countries like Bangladesh are trying hard to hang on to the low-wage jobs. This raises the question whether we have the right to deprive them of the opportunity to work.

Fashion is competing with other industries for valuable resources in terms of water, land, and fossil fuels. From an environmental perspective, reduced consumption of textiles would free up these resources to need more fundamental needs, like growing food or providing fuel for heating and cooking. However a reduction in the volume of production would have a huge impact on employment and growth across the world.

Another conflict lies between the desire to reuse garments to avoid further impact caused by producing more and the effect the large-scale export of second-hand clothing has had on indigenous textile production and markets in the developing world, in particular in Africa.

Changes to the textile industry are possible and necessary, but there are no simple solutions. Rather than radically changing the system with which textiles are produced and used, the improvement lies in making more ethical choices when choices can be made. For example, organic cotton can be grown after a transition period with yields marginally lower than nonorganic cotton. Working conditions can be improved by making sure that suppliers adhere to regulations and pay locally fair wages. Retailers can be encouraged to switch away from their least ethical suppliers.

Larger changes in the fashion and textile industry would require an underlying shift in the value that consumers give to clothing and textiles in the twenty-first century. Modern customers need to be reeducated to recognize quality in garments,

so that they realize once again when it is worth paying a higher price for particular garments. At the moment low prices have pushed both environmental exploitation and low wages. However, a greater understanding of what is worth paying for might entice people to pay more for better quality garments. Higher price points would allow for more sustainable forms of production and better treatment of workers. It would also give garments once again an enduring value of their own.

## Conclusions

Fashion and textile products are deeply personal and often have an emotional value for the consumer which far outweighs the monetary value of the garments. Looking at fashion and textile production and consumption from an industry perspective rather than the relationship of individuals to their garments, it is necessary to take a systemic perspective, which shows that fashion production and consumption share many steps with other consumer products.

The key values to consider in fashion and textiles are the sustainability of production, retail, use, and end of life and the ethical issues involved with the production of textile products in the developing world and the impact the large-scale export of secondhand garments to the developing world has on its indigenous industries.

The ability of the designers to make systemic changes in fashion and textiles is quite limited, unless the designers take charge of their own production and retail operations, which very few designers can. However designers need to be aware of the ethical and sustainability issues associated with their designs, so that they can make the least harmful choices, while influencing the overall system in a more favorable way. Rather than being able to outline a set of design principles or procedures to design for values, it is necessary in fashion and textiles to become aware of the choices and options designers have at each point. For example, the designers might like to use organic cotton, but cannot push for it at the price point they are working toward, but they will always have a choice between a number of options from which they can choose the best one. In this case this might mean sourcing cotton from a country where cotton production does not compromise other food crops, or choosing something that is dyed with less harmful chemicals.

Improving the textile and fashion life cycle requires systemic changes to the way textiles are made, designed, used, and disposed. A starting point would be greater awareness of the process by all stakeholders. For example, effective and potentially compulsory eco-labeling across the supply chain would make an enormous difference to the designers' ability to make informed decisions.

## Cross-References

► Design for the Value of Sustainability
► Design for Values in Economics

## References

Allwood J, Laursen S, de Rodriguez CM, Bocken N (2006) Well Dressed? The present and future sustainability of clothing and textiles in the UK. Institute for Manufacturing, Cambridge University, Cambridge

Black S (2008) Eco chic: the fashion paradox. Black Dog Publishing, London

Black S, Eckert CM (2009) Developing Considerate Design: meeting individual fashion and clothing needs within a framework of sustainability. In: Pillar F, Tseng M (eds) Making customer centricity work: advances in mass customisation and personalisation. World Scientific Press, Hackensack/London/Singapore, pp 108–146

Cappetta R, Cillo P, Ponti A (2006) Convergent designs in fine fashion: an evolutionary model for stylistic innovation. Res Policy 35(9):1273–1290

Crommentuijn-Marsh P, Eckert CM, Potter S (2010) Consumer Behaviour towards sustainability in Fashion. In: Keer10, Paris, Feb 2010

Davis F (1992) Fashion, culture, and identity. The University of Chicago Press, Chicago

Defra (2011) Sustainable clothing action plan. UK Government Department of Environment, Food and Rural Affairs, London

EC (2001) Green paper: promoting a European framework for corporate social responsibility. DG Employment and Social Affairs, European Commission, Brussels

Eckert CM, Stacey MK (2000) Sources of inspiration: a language of design. Des Stud 21 (5):523–538

Eckert CM (2001) The communication bottleneck in knitwear design: analysis and computing solutions. Comput Supported Coop Work 10(1):29–74

Eckert CM, Stacey MK (2001) Designing in the context of fashion – designing the fashion context designing in context. In: 5th design thinking research symposium, Delft University Press, Delft, 2001, pp 113–129

Eckert CM, Stacey MK (2014) Overconstrained and underconstrained creativity: changing the rhetric to negotiate the boundaries of design. In: Blessing LTM, Qureshi AJ, Gericke K (eds) The future of transdisciplinary design. Springer, London

Eskandarypur F, Black S, Eckert CM (2009) The development and positioning of the considerate design tool in the fashion and textile sector. In: Sustainable innovation 09, Farnham

Fletcher K (2008) Sustainable fashion and textiles. Earthscan, London

Flügel JC (1930) The psychology of clothes. Hogarth Press, London

IDE-JETRO (2013) http://www.oecd.org/dac/aft/AidforTrade_SectorStudy_Textiles.pdf

Kamal Y, Deegan C (2013) Corporate social and environment-related governance disclosure practices in the textile and garment industry: evidence from a developing country. Aust Account Rev 23(2):117–134

Lee M (2007) Eco-Chic: the Savvy shoppers guide to ethical fashion. Octopus Publishing Group, London

Lowe ED, Lowe JWG (1985) Quantitative analysis of women's dress. In: Solomon MR (ed) The psychology of fashion. Institute of Retail Management, New York

Nordestgaard M, Kirton-Darling J (2004) Corporate social responsibility within the European sectoral social dialogue. Transf Eur Rev Labour Res 10(3):433–451

OECD (2004) A new world map in textiles and clothing: adjusting to change. Organisation for Economic Co-operation and Development, Paris

Polhemus T, Procter L (1978) Fashion & anti-fashion: an anthropology of clothing and adornment. Thames and Hudson, London

Siegle L, Burke J (2014) We Are What We Wear: Unravelling fast fashion and the collapse of Rana Plaza, Guardian shorts, http://guardianshorts.co.uk/wearewhatwewear/

Simon-Miller F (1985) Commentary: signs and cycles in the fashion system. In: Solomon MR (ed) The psychology of fashion. Institute of Retail Management, New York

Sinha P, Shah R (2010) Creating a global vision for sustainable textiles. In: "Textiles: a global vision" the textiles institute centenary world conference, Manchester, 3–4 Nov 2010

Sinha P, Beverley KJ, Day CL, Tipi NS (2012) Supply chains for the management of post-consumer apparel waste: three scenarios addressing the UK-Tanzania context. In: Proceedings of the 18th international sustainable development research conference, The University of Hull, Hull

Sproles GB, Burns LD (1994) Changing appearances. Fairchild Publications, New York

Welters L, Lillethun A (eds) (2011) The fashion reader, 2nd edn. Berg, Oxford

WTO (2013) http://www.wto.org/english/res_e/statis_e/statis_e.htm

WRAP (2012) Valuing our clothes, report WRAP

WWF (2008) http://assets.wwf.org.uk/downloads/water_footprint_uk.pdf. Accessed 12 May 2011

Wynn D, Eckert CM, Clarkson PJ (2011) Simulating intertwined design processes that have similar structures: a case study of a small company that creates made-to-order fashion products. Int J Product Dev 14(1–4):118–146

# Design for Values in Healthcare Technology

Gert Jan van der Wilt, Rob Reuzel, and John Grin

## Contents



G.J. van der Wilt (✉) • R. Reuzel
Department for Health Evidence (133), Radboud University Medical Center, Nijmegen, The Netherlands
e-mail: gertjan.vanderwilt@radboudumc.nl; rob.reuzel@radboudumc.nl

J. Grin
Department of Political Sciences, University of Amsterdam, Amsterdam, The Netherlands
e-mail: j.grin@uva.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_36

**Abstract**

Communities struggle with finding ways for collaboratively exploring the value of healthcare technologies. Currently, a strong emphasis is being placed on the assessment of the costs associated with the health gains (expressed in quality-adjusted life years) that are achieved with these technologies. Following Hannah Arendt, we shall try to argue that such instrumental rationality is misplaced in discovering how technology can help to express human values. It typically reflects a society where processes of design and development, evaluation, and decision making involve separate trajectories and operate distinct from the realm of the lives of humans. We will present an alternative which is deliberative and transformative in nature. Its strengths and limitations will be explored, using the cochlear implant for deaf children as an example.

**Keywords**

Technology assessment • Values • Dialectic • Frame reconstruction • Evaluation as learning

## Introduction

Designing and creating technology is a fascinating process. It is fascinating because it reveals something about our physical environment (what is physically possible, for instance, will we, in fact, be able to live on Mars?), about our social environment (what is socially possible, for instance, will we be able to join forces to conduct research on human (im)mortality?), and about ourselves (how clever and ingenuous are we, for instance, will we be able to discover the theory of everything, if there is one?). It is also fascinating because it unites considerations of what is possible with considerations of what is desirable. This is neatly expressed in questions like "Would we clone them if we could?" (suggesting that there are probably too few Albert Einsteins in this world and definitely too few Marilyn Monroes). It is also fascinating because it testifies of a basic trait of modern societies, which is being dissatisfied with the world as it is (Heller 1999). However, mankind has not an impressive track record when it comes to presaging how technology will enable the creation of value. The realm of healthcare is no exception. Over the past decades, accumulating evidence has made it abundantly clear that healthcare technologies need not always be merely beneficial. Healthcare technologies that were once considered of great value turned out to be largely ineffective or positively harmful (Dutton 1988). In response to such experiences, we have witnessed a growing demand for the critical evaluation of healthcare technology, preferably at an early stage of development. This has resulted in a veritable healthcare evaluation industry, with its idiosyncratic standards and logic (Klein 1982). In this process, Health Technology Assessment (HTA) commissioning organizations such as NICE (the National Institute for

Health and Care Excellence) and HTA researchers have developed an evaluation methodology with a strong emphasis on cost-effectiveness. By translating health benefits into a single metric (quality-adjusted life years, or QALYs), league tables can be compiled, comparing different healthcare technologies in terms of costs per QALY gained. It is questionable, however, whether such analyses suffice to understand and communicate the value of healthcare technologies. Cost-effectiveness analysis is appropriate to explore whether a particular healthcare technology is a relatively efficient way of achieving certain preconceived goals. It is much less appropriate to collaboratively explore whether a technology results in practices that adequately reflect values that are considered important in the community concerned. The main objective of this chapter is to explore how existing methods of technology assessment may be adapted to better serve this role. The key features of this alternative approach are its interactive nature by engaging stakeholders, an emphasis on interpretation of evidence, and a focus on learning. We will refer to this approach as interactive technology assessment (iTA). In the next two sections, we will briefly discuss a number of key developments in the field of healthcare technology and the values that seem to be at work in this domain. We will then turn to the work of Hannah Arendt, who made a distinction between "work" (the construction of artifacts) and "action" (the public inquiry into the meaning of life) which is highly relevant to the question how technology bears on values and vice versa. We will then describe in more detail the method of iTA, followed by a case study, the application of this method to the evaluation of cochlear implants (CI) for deaf children. The chapter will be completed by a discussion of what we can learn from the case study: can iTA help stakeholders to discover how technologies should be designed and used in concrete situations so as to achieve maximal coherence among our multiple and varied value commitments, and, if so, what conditions should then be met?

## Trends in Healthcare Technology

Many healthcare technologies have grown out of ideas that became more serious in the course of the nineteenth century, have picked up speed after the Second World War, and continue to develop at accelerated speed at present. An example is the left ventricular assist device or LVAD, which developed from the heart-lung machine and which is currently developing further because of the emergence of novel materials, energy-saving designs, and improved methods of control of coagulation and inflammation. In the context of this contribution to the volume, it is impossible to provide anything near an overview of the wide and fast technological developments in healthcare. Suffice, perhaps, to note the closely intertwined development of GRIN; Genomics, Robotics, Informatics, and Nanotechnology. To give an example: developments in informatics and genetics have created unprecedented opportunities for identifying the genetic origin of many diseases. Functional analysis of these genetic deviations will pave the way for gene therapy. Currently, such

gene therapies are being developed for patients with serious eye disease and for patients with hearing disorders. Nanotechnology is part of so-called quadratherapeutics, where two types of nanoparticles are being combined with laser technology and radiotherapy in the experimental treatment of cancer. Another development that may have a significant impact on health is tissue engineering, allowing, for instance, the manufacturing of heart valves. The success of all these developments may be inferred from the estimated life expectancy of today's newborns, two third of whom is expected to live 100 years or more. Clearly, this will pose novel problems, for which novel solutions will be sought. Such success comes at a price, though: an ever-increasing part of our national budgets is being spent on healthcare.

## Design, Assessment, Evaluation

In the domain of healthcare, design may apply mainly to the development of devices, rather than to drug development. In drug development, large amounts of compounds are being tested in vitro and in animals for their therapeutic potential and safety profile. Design, here, may chiefly refer to the chemical alteration of such compounds in an attempt to improve their therapeutic-safety profile, to improvements in modes of administering drugs (as in the case of quadratherapeutics, mentioned above), or to attempts at targeting patients who are most likely to benefit and least likely to experience harm (personalized healthcare, such as restricting treatment with the antiretroviral drug Efavirenz to HIV patients with specific serotonergic polymorphisms). Not infrequently, drug discoveries are made serendipitously, as, for instance, in the case of Ritalin in the treatment of children with ADHD (Singh 2002). Device development usually extends over several decades, as, for instance, in the case of imaging devices (Blume 1992), the cardiac pacemaker, and the left ventricular assist device or LVAD (Sutton et al. 2007). The LVAD originated from the heart-lung machine, which enabled cardiac surgery. Since then, its development has been one of miniaturization, decreasing energy expenditure and improved control of thrombosis and infection, leading to various types of intracorporeal devices. Along the way, attempts at developing a total artificial heart were made but discontinued because of multiple problems that could not be resolved at the time. Until this day, unexpected problems have emerged with cardiac assist devices, demonstrating the strong "learning-by-doing" strategy in medical care (Starling et al. 2014). Cases like these point to the close link between designing, assessing, and evaluating healthcare technologies, with clinical experience giving rise to adjustments or technology redesign (e.g., in the case of the LVADs, continuous flow pumps as opposed to pulsatile flow pumps, development of new materials with different immunogenic properties, transcutaneous energy transfer, etc.). In this chapter, our focus will be on assessment and evaluation of healthcare technology, exploring what implications these may have for the design and redesign of such technologies.

## Values in the Domain of Healthcare

The values underlying healthcare are multiple and varied. Although far from exhaustive, the following values seem to be at play in the current development and provision of healthcare. Firstly, there is the preservation of life itself. If someone has sustained a myocardial infarction, clinical management is aimed at revascularization of the heart in an attempt to rescue the life of the patient. Similarly, if a newborn child suffers from life-threatening cardiac or respiratory failure, the child will be supported by a heart-lung machine. The same holds for patients who sustained severe neurotrauma, for patients with an acute appendicitis, etc.: many healthcare activities are aimed at preserving life in patients with life-threatening conditions. Another value is the relief of suffering. The suffering may be mainly physical, as in the case of patients with inoperable colon cancer, who are prescribed opioids to relieve intractable pain. The suffering may also be mainly mental in nature, as in patients with severe refractory depression who receive electroconvulsive therapy. Yet another value is the restoration of functions, enabling people to conduct valued activities. These would include cataract extraction to restore vision, and hip or knee replacement to restore mobility, although such interventions will also relieve the suffering (pain, depression) associated with these conditions. Still other healthcare activities are directed at the preservation or restoration of human dignity. An example of this would be the treatment of drooling in children with cognitive disability. Such treatment will not help to preserve life and is unlikely to relieve suffering. And yet, many of us would feel that it would testify of a lack of respect if drooling in such children would be left untreated, assuming that an effective treatment is available (which is, indeed, the case). Many of the healthcare activities, mentioned so far, could be conceived as being directed at lifting, alleviating, or compensating for the restrictions, imposed by disease. Or, positively defined, directed at preserving, restoring, or augmenting capability, as defined by Amartya Sen (Venkatapuram 2011). Quite another value, related to the delivery of individual healthcare, is respect for the patient's will. An elderly patient may, for instance, request deactivation of his or her cardiac pacemaker, believing that the pacemaker may prevent him or her from dying in a dignified way. Not granting such a request would testify of insufficient respect of a patient's autonomy, provided that the request was not a rash decision or inspired by a temporary and uncharacteristic loss of stamina. Another deeply ingrained value in the healthcare domain is the avoidance of inflicting harm to the patient ("primum non nocere", first do no harm). Perhaps, nowadays, the norm should read: avoid *unnecessary* harm. For, inflicting harm is unavoidably associated with any operative procedure and cancer treatment, be it surgery, radiotherapy, or chemotherapy. Still, inflicting avoidable or unnecessary harm to patients is considered a very grave offense, occasionally giving rise to the suspension or even prohibition of a professional's continued practice. Finally, there is increasing awareness that healthcare should not be wasteful, that there should be proportionality between effort, burden to the patient, and expected benefit, and that all patients deserve equal concern for their suffering and anxiety. These may be

considered as a commitment on the part of healthcare professionals to the value of justice or equity. To be sure, justice in healthcare has not been operationalized in a single, coherent fashion. On the one hand, there is a clear tendency toward maximizing the aggregate health gain by prioritizing the most cost-effective services. On the other hand, there is also the widely held conviction that healthcare should be used so as to mitigate gross inequalities in health experienced by humans during their life time. This conflict between a utilitarian and egalitarian conception of justice has been, and still is, one of the most vexing issues in healthcare policy (Daniels 2007).

## Hannah Arendt: A Typology of Human Activities

How, then, do the development of technology and the reflection on the value of the practices that emerge from the use of these technologies relate to each other? We think that Hannah Arendt's account of the human condition can be helpful in this respect. In *The Human Condition*, Arendt distinguishes three types of human activities: labor, work, and action (Arendt 1998). *Labor* involves activities (production and consumption) which are necessary for the maintenance of life. *Work* involves activities which are, literally, creative: the making of artifacts with which we equip the world and which serve to make our lives safer, more comfortable, more convenient, etc. *Action*, finally, involves communicative interaction between humans, aimed at collaboratively finding out the meaning and significance of their doings and beings. Each of these activities is, in its own way, indispensable to human life. In the absence of labor, we would not be able to continue living. In the absence of work, we would live, but in a very harsh way. In the absence of action, we might be able to live, and to live comfortably, but in a meaningless world.

Of the three types of human activities, action is the most elusive one, being not strictly necessary for the preservation of life, nor productive in the sense of resulting in tangible products that enrich our lives. And yet, it is quite clear that Arendt considers *action* the most crucially important activity, since it is through the collaborative construction of meaning that humans can achieve freedom and develop their identity. At the same time, however, she observes a tendency toward an ever greater emphasis on the preservation of life itself (*labor*) and the further improvement of the quality of our lives through the development of artifacts (*work*), at the expense of engaging in public deliberation about final ends (*action*). A clear indication of this is the tendency to use instrumental rationality, which is appropriate to *work*, as the sole guidance for assessing value. In our view, the strong focus on cost-utility analysis in the evaluation of healthcare technology is a striking example of this. However, as pointed out by Arendt, this is self-defeating. Having adopted instrumental rationality in the realm of *work*, thereby necessarily stripping everything of intrinsic value, humans cannot at the same time adopt this framework to ascribe – subjectively – meaning or value to artifacts, states of affairs, etc. Or, as Arendt puts it: Utility, when elevated to meaning, turns

into futility. The ascription of value and meaning is, according to Arendt, the prerogative of *action*, which is inherently deliberative.[1]

## The Meaning of Meaning

The key concept for *action*, then, is meaning, and in the context of assessing value of technology, the key question is not so much "What are the data?" but "What do the data mean, or what do they amount to?" This is not to say that data are unimportant. On the contrary. But the question of meaning goes *beyond* the data. In other words, when presented with data, we can intelligibly ask: What do these data mean? This question, about the correct interpretation of the data, refers to the practical implications: what, if anything, should we do and how should we act upon this information? The answer to that question cannot be obtained from the data themselves, but requires that the data are situated in a wider interpretative framework. Only then, we can make causal inferences and judge the relevance of the data, which is necessary to assess practical implications. To give an example, we may be presented with data on changes in global temperature. The question what these data mean goes beyond the data themselves. The question is about the possible and likely causes of these changes (causality), the likely future developments (prediction), and about the question whether we should be pleased about these changes, worried, or indifferent. Ultimately, it is about a judgment as to what an appropriate response would be. To take an example from the healthcare sector, we may be presented with data on the costs of healthcare over the past decades. Again, we can intelligibly ask: what do these data mean, what do they tell us, and how should we act upon them? Answers to these questions go beyond the data themselves and require an interpretative framework: What are the causes of the change in costs of our healthcare systems? Should we be pleased, worried, or indifferent? The change may be caused by the slow increase in productivity in healthcare as compared to other areas and interpreted as an indicator of economic growth. As such, at least from an economic perspective, the data should be a reason for joy, rather than concern, and give no cause to corrective measures (Baumol 1996). A final example to clarify the distinction between data and the interpretation of data might be changes in the burden of disease. Insel, for instance, reported that in spite of substantial efforts, the burden of disease associated with schizophrenia has not significantly decreased over the past decades (Insel 2010). The relevant questions are, again: what the reason for this might be, why it should be of our concern, and what our appropriate response to this finding might be? Only when we have answers to such questions, we can decide what we should do about it (if anything) and how we should act. "Action," as defined by Arendt, refers to this complex of activities, namely, acting on the basis of the purposeful and collaborative collection and interpretation of data. Or, as Farrell e.a. formulated it: "Assessment processes are

[1]For a comparable argument, see Richardson (2000).

embedded in different sorts of institutional settings, within which scientists, decision-makers, and advocates communicate to define relevant questions for analysis, mobilize certain kinds of experts and expertise, and interpret findings in particular ways" (Farrell et al. 2001).

## Plurality: It Ain't Necessarily So

Another central feature of Arendt's work, apart from the typology of human activities, is the essentially plural character of *action*. In other words, there are, and always will be, differences in how human beings interpret the world, including their own doings and beings. Our observations allow for multiple interpretations, that is, understandings of how things cohere and are brought about and judgments whether we should be concerned and how we should act upon them. This plurality follows from the fact that *action* requires that people are not forced to adopt a particular interpretation. Through *action*, people can achieve their freedom and, as such, reveal their identity to each other. This process results in, and at the same time requires, contrasting interpretations. Whereas in *labor* (activities that are necessary to maintain our lives) and *work* (activities that are necessary to make our lives comfortable, safe and efficient) the degrees of freedom are rather limited, this does not hold for *action*: "things ain't *necessarily* so." This plurality of interpretations gives rise to different views of what needs to be done and why: should we do something to counter global climate change, to bring the increase of public expenditures to healthcare to a halt, and to reduce the burden of disease associated with schizophrenia, and if so, what? Since these are clearly questions of collective, rather than merely individual action, some form of public deliberation is required, laying down procedural rules for who may be involved and in what way, and how, in the light of plurality, decisions are made as to what actions will be set in motion (Dryzek and Niemeyer 2010). In the next section, we will briefly discuss what this might look like in the context of the development and assessment of healthcare technology.

## Implications for the Development and Assessment of Healthcare Technology: Interactive Technology Assessment

Many people would perhaps readily concede to the distinction between the generation of data itself and their interpretation, usually referred to as the assessment – appraisal distinction. It is suggested, then, that assessment is an essentially value-free kind of scientific research, producing data for the decision-making process (appraisal), where values are brought to bear on the available evidence. This distinction is, however, utterly misconceived, since it fails to acknowledge that interpretative frames are already operative at the time of data collection. At that stage, they serve to define which data are considered relevant, plausibly associated with the intervention, and amenable to scientific research. If we do not recognize this role of interpretative frames in the production of

evidence, we fail to understand why the relevance of available evidence is not always endorsed by all stakeholders. To avoid this type of bias, then, involvement of the various stakeholders during the assessment process is vitally important (Reuzel et al. 1999). However, interpretative frames encroach even deeper in the process, since they also provide guidance to the design and development of healthcare technology itself (Schwarz and Thompson 1990). This would plead for the involvement of stakeholders at the stage of design and development of healthcare technology, in such a way that it can be qualified as a true instance of *action*. In the following sections, we will describe what such interactive technology assessment (iTA) and interactive technology design and development (iTDD) might look like. To do this, we will report our experience with the interactive evaluation of the cochlear implant for deaf children. This was an assessment of a technology which was more or less matured and which had to draw on the evidence that had been produced by then. The strengths and limitations of such an approach will be discussed, including a reflection on how the results of such deliberative practices may be integrated into processes where deliberation has been delegated to professionals (i.e., politicians).

## Case Study: Interactive Evaluation of Cochlear Implants for Deaf Children

In this section, we will present our case study, the cochlear implant for deaf children. We will start by briefly describing the societal debate regarding this technology. This will be followed by a description of our interactive evaluation of the technology in terms of method and its key results.

### The Technology and the Opposition from Deaf Communities

A cochlear implant (CI) is a device that is meant as a complete substitute for the function of the outer, middle, and inner human ear (O'Donoghue 2013). It consists of a sound receiver, a processor, and an electrode which transmits electrical signals to the acoustic nerve. It is used in adults and children with profound sensory-neural hearing loss. Especially when surgeons started to use the technology in deaf children, this provoked a fierce debate, with strong opposition from organizations advocating deaf culture (Hyde and Power 2006). Whereas proponents of CI considered the technology as an unprecedented opportunity for deaf people to integrate into hearing society, many deaf organizations across the world remain critical unto this day. One of the concerns was the publicity about the cochlear implant. This publicity was considered as one suggesting that

> Deaf people are ill or incomplete individuals, are lonely and unhappy, cannot communicate effectively with others, and are all desperately searching for a cure for their condition. Such publicity is highly inaccurate. It also demeans Deaf people, belittles their culture and language, and makes no acknowledgment of the diversity of lives Deaf people lead, or

> their many achievements. It is stressful for hearing parents of deaf children who are already struggling to come to terms with their child's deafness and are given a false impression that the implant will 'cure' their child.. (Deaf Australia, Policy on cochlear implants; www.deafau.org.au/info/policy_cochlear.php Accessed on January 29th, 2014)

With respect to the use of cochlear implants in deaf children, great concern has been expressed:

> The decision to implant such children is usually made by parents and guardians and Deaf Australia has long believed that parents and guardians do not have access to full and accurate information about the implication of deafness for their children's lives. Parents are usually led to see their children as pathologically deficient and little information is available to them about the history, culture and language of Deaf people, or the possible lives of Deaf people in our society. Until such time as more complete information is available for parents, and more productive associations develop between parents of deaf children and adult Deaf people, Deaf Australia feels that decisions to implant young deaf children are questionable.. (Deaf Australia, Policy on cochlear implants; www.deafau.org.au/info/policy_cochlear.php Accessed on January 29th, 2014)

## Interactive Evaluation of CI for Deaf Children: Methodology

The interactive evaluation which we conducted on CI for children was based on the method for joint evaluative design, described by Guba and Lincoln (Guba and Lincoln 1989). Briefly, it consists of identifying and engaging all stakeholders, where stakeholders are defined as those people or groups of people who are likely to experience the consequences of the evaluation and inquire into their constructions of problem and judgments of potential solutions, thus gradually developing a so-called joint construction of the problem and a preferred solution as the final result of the evaluation. Generally speaking, in the case of healthcare technology, stakeholders would typically involve patients and their relatives, (allied) healthcare professionals, healthcare insurers, manufacturers, hospital managers, representatives of advocacy groups, and policy makers. These may be approached directly by the members of the evaluation team, may come forward of their own accord, or become involved through snowballing. In the case of the evaluation of CI, 51 stakeholders participated, including ENT surgeons, audiologists, teachers from Institutes for deaf children, social workers, speech therapists, policy makers, hospital managers, and representatives from various advocacy groups (Reuzel 2002). Each of these participants was interviewed individually. The purpose of the interview is to reconstruct the interpretative frame of the stakeholder. An interpretative frame is the ensemble of a stakeholder's judgment of various solutions to a specific problem, problem definition, background theory, and values (Schön and Rein 1994; Grin and van de Graaf 1996). Typically, an interview starts by asking what the interviewee considers as particularly problematic about the current situation (problem definition). From there, the interview may proceed by asking what strategies the interviewee considers most likely to resolve the problem (judgments of solutions). To better understand *why* a stakeholder considers a particular situation problematic

and *why* he or she considers particular strategies most likely to be effective in resolving the problem, it is important to reconstruct background theory and values. The method of frame reconstruction (Grin et al. 1997) has been applied to a wide number of topics (Fischer et al. 1993, 2012) and has been found a valid and reliable method of qualitative data analysis (Moret-Hartman et al. 2007a). After checking and, if necessary, revising the reconstructed frame (respondent validation), a second interview round is held, where interviewees are confronted with the reconstructed frames of all participants (anonymously) and invited to respond to them. In particular, interviewees are asked whether, based on the reconstructed frames of the other participants, they wish to revise (parts of) their own interpretative frame and whether there are particular issues that they would like to have investigated or checked in the literature. In doing so, the material particularities of the situation are explicitly considered. The evaluation team, then, explores such issues by reviewing the literature and consulting experts. The findings are then fed back to the participants in a third interview round, where interviewees are also asked to review their own interpretative frame, taking into account the interpretative frames of the other participants and any findings from the literature or expert consultation. They are also invited to attend a meeting with all other participants, where the key findings will be presented and unresolved or contentious issues will be discussed, aiming at an agreement on how to proceed (or not) with the technology under investigation.

## Interactive Evaluation of CI for Deaf Children: Key Findings

Here, we will report the key findings of the interactive evaluation of CI for deaf children that intended to design the socio-technical practices around CI so as to incorporate concerns from the stakeholders involved; a full account is presented elsewhere (Reuzel 2002). Firstly, it was found that those who were in favor of CI differed in their problem definition from those who were more critical of the technology. To the former, the key problem was simply that deaf children cannot hear, and that, for this reason, deaf children are unlikely to develop well in a hearing society. A CI is a logical remedy to this problem, although it is admitted that it is still an imperfect technology. Those who were more critical of CI framed the problem in a slightly – but importantly - different way. For them, the key problem is that deaf children, during the first months of their life, do not receive the sort of sensory input that is appropriate and necessary for their social, emotional, and particularly cognitive development. For them, the consistent use of sign language by the parents in the communication with their child is critical to the child's development.

Interestingly, these positions turned out to be associated with differences in background theory and values. Proponents of the CI appeared to be reasoning from the assumption that sign language and spoken language are mutually competitive: acquisition in linguistic competence in sign language will interfere with acquisition in spoken language and vice versa. This explains their negative attitude

toward the acquisition of sign language by deaf children. Those who were more critical of CI appeared to reason from the assumption that the two linguistic modalities could, in fact, be mutually reinforcing, as is known from bilingualism in spoken language. In terms of values, the views of proponents of CI could be largely resolved to the open future argument: do not choose options that are likely to constrain the range of future opportunities that are open to the developing child. Those who were more critical of CI were more likely to emphasize the importance of cultural diversity: there are multiple ways of organizing social life that can serve as a basis for the development of mutual and self-respect. Further key differences were observed in the appreciation of deaf culture: is it, or is it not, a fully fledged culture, and, relatedly, is sign language a fully fledged language? Finally, proponents and opponents of CI differed in their view of the significance of deaf culture to the social and emotional development of deaf children, notably with respect to the development of their self-esteem. It is important to note that most participants were not aware of the fact that they were reasoning from these premises, and how they influenced their problem definition and appreciation of the technology. Nor were they always able to provide compelling evidence or arguments in support of these premises. For this reason, we tried to retrieve literature that might shed more light on the issue of the antagonistic or synergistic nature of the relation between development in spoken language and sign language. This produced data from studies that had been conducted in Sweden, where deaf children had been followed over time, from before cochlear implantation to several years thereafter (Preisler et al. 2005). The researchers had used both, quantitative and qualitative methods, observing the children in their own environment, and investigated various aspects of their communicative development. Children who were most proficient in sign language at the time of cochlear implantation were more likely to benefit from the CI in terms of developing proficiency in spoken language. Although certainly not conclusive, these data seem to suggest that the two linguistic modalities can be synergistic, rather than antagonistic.

At the final meeting that was organized, it was agreed that early identification of deafness in newborn children is of crucial importance and that deaf children probably benefit from the consistent use of sign language by their parents, irrespective of whether they will receive a cochlear implant or not. Clearly, such recommendation has considerable implications for the parents, for how they are informed about sign language, deaf culture, and CI, and for creating sufficient training capacity in sign language.

In summary, then, the interactive evaluation had helped to overcome the impasse, following the intense debate on cochlear implants for deaf children, and it helped to develop a practical solution that was acceptable to all participants. It is important to note that the emphasis in an interactive approach is on *learning* between actors with different frames (Grin and van de Graaf 1996) and that it is both *summative* (what is the value of the technology, when used in this way?) and *formative* (how can the value of the technology be improved through the specific way in which the technology, and associate practices, is (re)designed?) (Scriven 1996).

## Frame Reflective Analysis: A Reflection

The aim of our contribution to this handbook was to explore what can be done to gain a better understanding of how technologies relate to values that are considered important in our society and how we can find out what implications these value commitments might or perhaps should have for the choices that are made in the context of technology design and development. We have drawn on our experience with evaluating CI for deaf children, where we used a method of TA which is interactive, which comprises reconstruction of stakeholders' interpretative frames, and which is directed toward learning. In the remainder of this chapter, we will discuss what might be learned from this case study. By relating it to a number of insights from various disciplines (policy sciences, philosophy, social science of technology), we will try to clarify its key characteristics and prerequisites.

## The Dialectic Nature of iTA

As we have seen in the case of the cochlear implant, interactive evaluation is primarily concerned with the practical interpretation of facts: what implications should our findings have for the design of a social-technical innovation? In the case of the cochlear implant, we may consider evidence of how deaf people live in hearing societies. The question of interpretation goes beyond such data and asks: what are the underlying causes, how do we judge the situation from a normative perspective, and, consequently, what would be an appropriate response to the current situation? We have seen that in the case of CI, the available evidence was interpreted quite differently by different stakeholders, resulting in quite different views of what needs to be done. iTA may be thought of as an attempt to bring about a synthesis between actors with different interpretative frameworks. As such, it may be considered as truly dialectic: a discourse between multiple stakeholders who hold different views on a particular subject, who wish to reach agreement on the appropriate course of action on the basis of reasoned arguments. It will, therefore, be beneficial especially in situations like this, where actors' problem understandings seem incompatible, i.e., when, at first sight at least, these problem definitions do not allow for a common solution. In such cases, it is quite likely that so-called Type III errors (solving the wrong problem) are made – more precisely, that the problem is defined in a way that privileges the viewpoint of particular actors. As Hoppe argues, this often is not an analytical mistake, but a matter of contextually prevailing rationality and network structures (Hoppe 2010). What we saw in the CI case was that stakeholder engagement and frame reconstruction resulted in the development and exploration of alternative frames, allowing for discovering what was "behind" proposed solutions (the CI). As such, the approach resulted in greater plurality in interpretation of the situation. On the basis of this feature, iTA may be conceived as a practical attempt to foster *action*, as defined by Arendt.

It is important to realize that deciding whether problem definitions are incompatible is not merely a matter of fact. Even where this seems to be the case, it may

be a matter of creative judgment. Basically, this is because problems and conceivable solutions themselves, rather than being objectively given, are constructed through judgment. The implication is that the conclusion that no joint solution for a problem exists is contingent upon preceding judgment, or lack of prudent judgment, and may change through a process of further, creative judgment.

## The Central Role of Judgment

The CI case also revealed the central role of judgment: judgment of the situation (what problems are experienced by deaf children growing up in hearing societies?), judgment of proposed explanations (do these problems originate from the lack of hearing in deaf children?), judgment of proposed solutions (e.g., the CI), judgment of available evidence (e.g., what can we infer from the available evidence concerning the clinical and cost-effectiveness of CI?), etc. Judgment here refers to the Aristotelian notion of *phronèsis*, which has been key to the development of interactive technology assessment. *Phronèsis* is practical wisdom: knowing what needs to be done in concrete situations in order to achieve maximal coherence among our multiple and varied value commitments (Richardson 1994). By its nature, a key feature of *phronèsis* is interpretation: it involves a double hermeneutic of interpreting both generic principles and the particularities of the situation. Following Arendt, it also essentially involves that different viewpoints are synthesized. This notion of *phronèsis* has been taken up since the mid-1980s in a then emerging approach within the policy sciences, sailing under the flag of the "argumentative turn" (Fischer et al. 1993, 2012) or "interpretative policy analysis" (Yanow 2000). Emphasizing the hermeneutic nature of this process, *phronèsis* has been conceived as a notion that may help understand the *formation* of interests, preferences, problem definitions and policy options, bringing back in the essence of politics, and human (inter)action more generally where it had been discarded by rational choice theory's assumption that these are merely exogenously given and thus constant (Hindmoor 2006). Obviously, such understanding is especially useful when analyzing socio-technical change, in which preferences may evolve both as part of the (i.e., as a driver of or response to) innovation. Interestingly, this is exactly what happened in the case of the iTA of CI: stakeholders held different views on the technology; through frame reconstruction these could be resolved into different background theories; this occasioned partial revision ("learning"), resulting in agreement on how to proceed ("irrespective of whether a child will be notified for CI, start using sign language as early and consistently as possible"). In Richardson's (1994) terms, it is an instance of the revisability of goals.

In an insightful review of this literature, Loeber has outlined the main features of this rich understanding of *phronèsis*: it concerns the collective and is thus not necessarily a private virtue; it reflects the interface between the reflective and the practical order; and it is principally a transformative capacity, oriented to designing

both action and transformation (Loeber 2007). Drawing on philosophical hermeneutics and pragmatism, she then argues that these elements imply that knowledge production:

(i) Should be organized interactively and participatory, enabling deliberative exchange of viewpoints
(ii) Should adopt an interpretative view on data gathering (which qualifies, but certainly not precludes, empirical-analytical analysis)
(iii) Should be "set up to induce transformative learning on the part of the participants as a result of the interactions with, (i) the problem situation, (ii) the way in which others define the problem and (iii) the particularities of the contexts in which these others operate. [It should enable] reflection on participants' own interpretive frames, allows for a 'fusion of horizons' and potentially enables participants to act in line with the newly required understanding" (Loeber 2007, p. 399).

This summarizes, in a way which is equally precise as concise, the rationale for, and fundamental nature of, iTA, which thus transcends the mere idea of "participatory" policy analysis: it enriches the latter with elements of critical as well as hermeneutic-interpretative analysis, into what Hoppe has called "forensic policy analysis" (Hoppe 2010). iTA thus becomes a form of analysis which not only assesses but also shapes the design of socio-technical innovations and policies, which does so through considering them in the light of (different views on) the good life and individual identities, and which, in doing so, helps actors to articulate, and deliberate on, their different viewpoints on the predicament in which we find ourselves and actions that seem most appropriate to turn what is into what ought to be.

## iTA: Contexts of Application

Thus understood, interactive technology assessment as a form of deliberative analysis and design may yield at least two contributions. First, in the context of health policy making, iTA as a form of deliberative policy analysis could inform policy making and thus help avoid Type III errors (the solving of the wrong problem). While this might help to embed a novel health intervention in health practices so as to do more justice to the life world of patients, opportunities to adapt the technology are likely to be limited in that stage. Thus, a second context of application where iTA might be useful are the practices in which medical innovations are being developed.

The rationale for using iTA to support formal policy making would be that it offers a departure from the often encountered practice that policy solutions reflect merely the problem definitions of those actors who have privileged access to the policy process. In healthcare these tend to be actors on the supply side, as the key policy mechanisms indicate "evidence-based admission" and "reimbursement" of drugs and devices, produced by an innovation system in the form of a tightly knit

network of public knowledge institutions, the medical profession and industry. This intertwined policy and innovation network forms a typical case of a closed, institutionalized network, hardly accessible to other parties. Policy making is typically information driven, drawing on a community of experts with shared knowledge base and standards of rationality. In case of the health system, this system represents the dominant medical rationality (Schuitmaker 2013), understanding health and disease as located in a (universal) body and thus as mainly determined by healthcare interventions, ignoring other determinants like lifestyle, life circumstances, and agency of patients (Wilt 1995).

As Hoppe (Hoppe 2010) points out, such networks are well able to deal with structured problems, i.e., problems on which a fair degree of normative consensus and factual certainty are presumed to exist. Or, more accurately, such networks tend to treat all problems as structured ones – more specifically as problems structured in the same way as earlier ones – and shape solutions accordingly. This has many potential drawbacks. Health problems may be suboptimally addressed, especially when they differ in nature from the problems that used to dominate the agenda. One important example concerns the now increasingly important class of noncommunicable diseases (including diabetes, COPD, obesity, and cardiovascular problems) where precisely these other three determinants may play key roles, as we have shown for, e.g., irritable bowel syndrome (Moret-Hartman et al. 2007b) and low back pain (Gielen and Grin 2010). The cochlear implant example is another case in point, which moreover suggests that such one-sided rationality may also lead to innovations being rejected by those for (!) whom they were developed, generating more societal controversy and less health effects than desired.

Under such circumstances, iTA as a deliberative, participatory, transformative design practice may help restructure the policy problem, provided that the conditions, defined by Loeber, are met. The cochlear implant case shows that it may be difficult to meet all of these conditions. The first condition, organizing the analysis in an interactive, participatory and deliberative fashion, was largely met. Participants were selected to include technology developers (as a manufacturer), those providing or applying it (e.g., as a physician), users or receivers of the technology, decision makers, and those supporting or helping patients in using the technology (e.g., a nurse, a social worker, a teacher and so forth). Also, participants could suggest to invite others; newspaper advertisements were placed in national papers; a web page was issued; and it remained possible for additional participants to be included at later stages of the process.

Meeting the second criterion suggested by Loeber, adopting an interpretative, rather than a technical, approach to data gathering, proved to be more difficult. It was difficult for the evaluators not to reproduce – either by habit and training – or to accommodate institutional requirements or credibly respond to interventions by health professionals (Reuzel et al. 2007). At least as problematic was that the available literature was so strongly dominated by established medical rationality that it was difficult to test and/or underpin other viewpoints (an important exception being the longitudinal studies conducted by Preisler and her colleagues, as necessary for restructuring the problem) (Preisler et al. 2005).

The third condition mentioned by Loeber, inducing transformative learning (a learning process giving rise to novel, unanticipated interpretation of the issue and novel approaches to its resolution) was both promoted by the just mentioned rules for deliberation and participation, as well as by a project design favoring learning through successive interview rounds (Guba and Lincoln 1989) *and* hampered by the imbalances in terms of underlying knowledge base between various viewpoints. Other hampering factors included a lack of trust and power imbalances between participants (Reuzel et al. 2007).

This takes us to another set of conditions, pertaining to the relationship between the iTA practice as a locus for "problem restructuring" and the closed, institutionalized network with its tendency to deal with the problem through "normal" problem structuring. In the case of the cochlear implant, iTA was used to increase the acceptability of the technology by shaping its conditions of application. This was done in a context in which the decision to introduce (i.e., admit and reimburse) CI in the Netherlands had already been taken, while the fact that the technology already had been developed and shaped without much involvement of the deaf and their relatives had eroded acceptance and trust between them, the medical establishment, and government. This further added to power differences between participants that already were there due to differential access to decision making and the dominance of medical rationality. It would be crucially important if the policy community would learn from such cases so as to introduce institutionalized warrants to nurture such deliberative exercise and adequately absorb their details.

Finally, the agency of the evaluator seems of critical importance in establishing and maintaining the conditions for deliberative, participatory, and transformative judgment in the iTA. Thus, developing a critical mass of evaluators with the requisite competences is of key importance. In addition to process competence and strategic insights, sufficient mastering of, and a capacity to critically scrutinize, medical knowledge are quintessential in this respect.

## Implications for Design for Values in Healthcare

What the case of pediatric cochlear implantation has shown is how the life world of intended users is easily obscured from sight by R&D and decision-making procedures in force that have a political rationality. To be sure, CI has been developed on the basis of a vision of this life world. For example, CI was originally designed as a safety tool that helped to pick up sound and hence warn for traffic and other dangers, rather than a communication device. When, soon after that, CI turned out to provide with a better hearing than had been anticipated, a new vision of deaf people being able to participate in a hearing society developed, giving direction to the process of further design. However, it is questionable whether the deaf themselves were in a position to contribute to shaping this vision, particularly when the vision on deafness as a handicap became the dominant paradigm in R&D circles. As soon as this was a political reality, a fierce debate ensued about the

validity of the paradigm. As all stakeholders tried to "win" this debate, it appeared as if the life world was buried under a pile of arguments for or against CI.

What the case of pediatric cochlear implantation has also shown is that behind the roles stakeholders play in this political arena are humans who at a deeper, normative level have more in common than can surface in such an arena. Fights over funding CI at the expense of support for the deaf community and the acknowledgement of sign language, for example, had obscured the fact that all persons involved sought to achieve happiness for deaf children. In their life world, happiness involved: feeling safe in a trusted social environment, being able to communicate with persons in this environment, and having the opportunity to pursue what you value in life. Based on these shared values, cochlear implantation could be redesigned and accepted as a social construct, even though it did not change fundamentally as a device.

It is here that iTA demonstrates its value: its potential to have persons involved leave the political arena and share their life worlds instead. In these life worlds, there is nothing to win and much to gain.

## iTA to Shape R&D Programmes

The example of the cochlear implant suggests that an interactive approach might also create a different relationship between R&D programmes and design priorities and patients' needs and desires: witness the examples where much more congruency could be achieved than anticipated by those involved. Given that path dependencies (limitation of options as the result of previous decisions) are likely to be involved in technological development, there might have been more room for accommodating these needs and desires, had they been taken into account at an earlier stage, when development was still relatively open. It could, in other words, help pre-empt Type III errors (solving the wrong problem) in the process of designing the innovation that is particularly prone to such errors as the main players – technology developers and doctors – share by and large the same medical rationality.

By going through such a process at an early stage of development, the technology and the features of its application practices may be designed so as to reflect much better the life world, needs and demands of recipients, which in turn may yield timely recognition of important conditions for application, wider acceptance (and hence application), and more health effects. By repeating part of the process after some experiences have been gained, further development may be oriented toward improvement of these concerns.

In the past few years, various approaches to involving patients and other considerations early in the process have been developed. One example is the ELSI approach to elucidate ethical, legal, and social implications of a technology under development (see, for instance, www.who.int/genomics/elsi). While this approach may help to timely anticipate moral, social, and legal issues, it pays much less attention to the feedback of such considerations into the development and design process. The latter aspect gains much more attention in a second

approach, constructive technology assessment, or CTA (Schot 1992), incorporating understanding of innovation dynamics in the form of such notions as "innovation journeys," "development trajectories," etc. While iTA, when used in innovation processes, could benefit from these insights, it could add to CTA the notion of phronèsis, yielding a better understanding of how novel considerations could enter the process through creative, communicative judgment, taking into the account the discursive and material particularities of the situation. We would suggest to coin such an approach interactive technology design and development or iTDD.

Now turning to the practice and methodology of TA in such a context, we may be brief on the rules that should prevail. Given the above rationale, an orientation toward the collective, a capacity for representative thinking, and a keen eye for the life world of the various stakeholders are clearly indicated. Thus the rules should, again, focus on warranting the deliberative, participatory, transformative nature of technology assessment.

One crucial issue would be to determine in what cases to undertake such an iTDD. While generic prescriptions seem hardly possible, one circumstance under which more often than not doing a constructive iTA could make a difference is in the early stage of a non-incremental innovation, like CI, or the more recent joint applications of genomics, robotics, informatics, and nanotechnology (GRIN). Such an approach may also help resolve controversies (see again the CI case) or help where compliance appears problematic. Crucial and particular examples where the latter is important concern interventions that presuppose, or seek to promote, agency by the patient and/or her informal caretakers. With the increasing recognition of lifestyle as a health determinant, the increasing interest in prevention, and the growing emphasis on more patient agency in healthcare, there seems to be a reason to develop a better capacity for such forms of evaluation.

A final issue for the practice of such constructive iTDD concerns how to actually achieve phronèsis against the dominance of one particular rationality. Practices of designing medical innovation are likely to be characterized by what Flyvbjerg (1998) has called the "rationality of power," a notion that refers both to the fact that a practice tends to be dominated by the rationality of the most influential actors and the fact that those who "possess" rationality are likely to have the power to impose their problem definition. To be sure, we are not necessarily, and certainly not primarily, referring to attempts to deliberately control others. What we are hinting at are more subtle mechanisms: for instance, that it is through the minds and hands of innovators that concerns of others are translated into different design choices or that designers tend to orient themselves on the practices where the innovation is to be realized and where most likely established medical rationality is to prevail. One thing that may help here is the participation of intermediaries who are able to translate the concerns and life world realities of patients to medical innovation design practices. Nurses and other paramedics, as well as patients with a (para)medical training, may make important contributions here. Thus, stakeholders may be conceived as parties who are constantly trying to further elaborate and test their interpretative frames. It is not difficult to see that some parties have and have had more opportunities to do so than others. iTA and iTDD can be seen as means to

redress such inequality, in the common interest of creating a greater plurality in interpretation of our life world and in our repertoire for responding to this predicament (Schwarz and Thompson 1990).

## Conclusion

The design and development of technology is a fascinating process, fascinating because it reveals the impressive capacity and achievement of mankind on the one hand and his limitations in developing a proper understanding of these activities on the other hand. Specifically, we seem to have difficulty in reaching a public understanding of the value of technologies (in Arendt's words: action vs. work). Technology assessment has been devised to fulfill this role. However, particularly in the realm of Health Technology Assessment, strong emphasis has been placed on various types of risk-cost-benefit assessment. These provide powerful tools to assess whether specific technologies represent efficient and safe ways of achieving particular ends. They are, however, insufficient in supporting the value inquiry, characteristic of *action*. We believe that stakeholder engagement in combination with reconstruction of their interpretative frames renders TA much more relevant to this task. This is not trivial since, following Arendt, it is in *action* that humans are capable of developing their identity and of exercising their capacity for freedom. Last but not least, as a spin-off, health professionals may gain a more satisfactory relationship with their clients in this way.

It is in this vein that we seek to reconcile labor, work, and action and thus retain the relation between the *vita activa* as a whole and the betterment of the human condition. For technologies like cochlear implantations are products of work that are meant to improve the human condition, both from an individual and a societal perspective. Yet, they also show that when work and action get out of sync and communication between various actors falls short, intentions for the good easily turn out to be courses to the bad. Through deliberative, participative, and creative procedures, a broader perspective on the human condition may be retained, embedding work and action as aspects of a life world that should not be separated.

It is to Richard Feynman that the observation is attributed that "Philosophy of science is about as useful to scientists as ornithology is to birds." Not surprisingly, perhaps, we disagree. It is through the work of Hannah Arendt and many others that we can come to see more clearly the potential value of technology assessment and why the current practice of technology assessment does not live up to this expectation.

## Cross-References


- ▶ Design for the Value of Human Well-Being
- ▶ Design for the Value of Inclusiveness
- ▶ Human Capabilities in Design for Values
- ▶ Participatory Design and Design for Values
- ▶ Technology Assessment and Design for Values

## References


Arendt H (1998) The human condition, 2nd edn. Chicago University Press, Chicago

Baumol W (1996) Children of performing arts, the economic dilemma: the climbing costs of healthcare and education. J Cult Econ 20(3):183–206

Blume S (1992) Insight and industry. On the dynamics of technological change in medicine. MIT Press, Cambridge, MA

Daniels N (2007) Just health. Meeting needs fairly. Cambridge University Press, Cambridge

Dryzek JS, Niemeyer P (2010) Foundations and frontiers of deliberative governance. Oxford University Press, Oxford

Dutton DB (1988) Worse than the disease. Pitfalls of medical progress. Cambridge University Press, Cambridge

Farrell A, Vandeveer S, Jager J (2001) Environmental assessments: four under-appreciated elements of design. Glob Environ Chang 11:311–333

Fischer F, Forester J (eds) (1993) The argumentative turn in policy analysis and planning. Duke University Press and University College London Press, Durham

Fischer F, Gottweis H (eds) (2012) The argumentative turn revisited: public policy as communicative practice. Duke University Press, Durham

Flyvbjerg B (1998) Rationality and power. Democracy in practice. University of Chicago Press, Chicago

Gielen AJ, Grin J (2010) De betekenissen van "evidence based handelen" en de aard van "evidence". Lessen rond rugscholen en radicalisering. Hoofdstuk 2 (p. 59–78). In: Verlet D, Devos C (red.). Efficiëntie en effectiviteit van de publieke sector in de weegschaal. Studiedienst van de Vlaamse Regering, Brussel

Grin J, van de Graaf H (1996) Technology assessment as learning. Sci Technol Hum Val 21 (1):72–99

Grin J, van de Graaf H, Hoppe R (1997) Interactive technology assessment: a guide. Rathenau Institute Report W57. The Hague, SDU. (in Dutch)

Guba EG, Lincoln YS (1989) Fourth generation evaluation. Sage, Newbury Pak

Heller A (1999) A theory of modernity. Wiley-Blackwell, Hoboken

Hindmoor A (2006) Rational choice. Palgrave, Basingstoke

Hoppe R (2010) The governance of problems. The Policy Press, Bristol, p. 132 ff

Hyde M, Power D (2006) Some ethical dimensions of cochlear implantation for deaf children and their families. J Deaf Stud Deaf Educ 11(1):102–111

Insel TR (2010) Rethinking schizophrenia. Nature 468:187–193

Klein R (1982) Performance, evaluation, and the NHS: a case study in conceptual perplexity and organizational complexity. Public Adm 60(4):385–407

Loeber A (2007) Designing for Phronèsis: experiences with transformative learning on sustainable development. Crit Policy Anal 1(4):389–414

Moret-Hartman M, Reuzel R, van der Wilt GJ et al (2007a) Validity and reliability of qualitative data analysis: inter-observer agreement in reconstructing interpretative frames. Field Methods 19:24–39

Moret-Hartman M, van der Wilt GJ, Grin J (2007b) Health technology assessment and ill-structured problems: a case study concerning the drug mebeverine. Int J Technol Assess Health Care 23(03):316–323

O'Donoghue G (2013) Cochlear implants – science, serendipity, and success. N Engl J Med 369:1190–1193

Preisler G, Tvingsedt AL, Ahlstrom M (2005) Interviews with deaf children about their experiences using cochlear implants. Am Ann Deaf 150(3):260–267

Reuzel RPB (2002) Health technology assessment and interactive evaluation: different perspectives. PhD Dissertation, Radboud University

Reuzel RPB, van der Wilt GJ, ten Have HAMJ et al (1999) Reducing normative bias in health technology assessment: interactive evaluation and casuistry. Med Health Care Philos 2(3):255–263

Reuzel R, Grin J, Akkerman T (2007) Shaping power, trust and deliberation: the role of the evaluator in an interactive evaluation of cochlear implantation. Int J Foresight Innov Policy 3 (1):76–94
Richardson HS (1994) Practical reasoning about final ends. Cambridge University Press, Cambridge
Richardson HS (2000) The stupidity of the cost-benefit standard. J Legal Stud 29(2):971–1003
Schön DA, Rein M (1994) Frame reflection. Towards the resolution of intractable policy controversies. Basic Books, New York
Schot JW (1992) Constructive technology assessment and technology dynamics: the case of clean technologies. Sci Technol Hum Val 17(1):36–56
Schuitmaker TJ (2013) Persistent problems in the Dutch health care system: learning from novel practices for a transition in health care with the UPP framework. PhD dissertation, University of Amsterdam
Schwarz M, Thompson M (1990) Divided we stand. Re-defining politics, technology, and social choice. Univ Pennsylvania Press, Philadelphia
Scriven M (1996) Types of evaluation and types of evaluator. Am J Eval 17(2):151–161
Singh I (2002) Bad boys, Good mothers, and the "Miracle" of Ritalin. Sci Context 15(4):577–603
Starling RC et al (2014) Unexpected abrupt increase in left ventricular assist device thrombosis. N Engl J Med 370:33–40
Sutton R et al (2007) History of electrical therapy for the heart. Eur Heart J 9(Suppl I):I3–I10
van der Wilt GJ (1995) Alternative ways of framing Parkinson's disease: implications for priorities for health care and biomedical research. Ind Environ Crisis Q 9(1):13–48
Venkatapuram S (2011) Health justice. Polity Press, Cambridge
Yanow D (2000) Conducting interpretative policy analysis. Sage, Thousand Oaks

# Design for Values in ICT

Alina Huldtgren

## Contents



A. Huldtgren (✉)
Fachhochschule Düsseldorf, Düsseldorf, Germany
e-mail: alina.huldtgren@fh-duesseldorf.de



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_35

**Abstract**

Information and communication technologies (ICT) are becoming pervasive. ICT development has accelerated, and within a few decades its use has expanded from particular work domains to diverse areas of work and everyday life. Consequently, the range of ICT stakeholders expanded from highly trained experts to all kinds of people with varying expertise and abilities. Sometimes, even people, who are not active users, are affected by the surrounding ICT.

Since ICT influences stakeholders' lives and in particular also their values, the ethical impact of ICT and the active consideration of values throughout design of ICT have become topics for research in several disciplines, including among others computer ethics, social informatics, or human-computer interaction. This chapter provides an overview of the history of ICT; different approaches to investigating, analyzing, and incorporating values in ICT; and practical methods to account for values in the ICT design process.

**Keywords**

Information and communication technologies • Value-sensitive design • Pervasiveness • Emerging technologies • Design methods

## Introduction

The fast-paced development of information and communication technology (ICT) has led to its widespread use in many domains. Research in the field has shifted focus in the last decades from purely technical advancement towards socio-technical aspects including the social and cultural context of ICT. With the advance of the Internet, eCommerce, Web 2.0, and recent ubiquitous computing trends, ethical concerns about privacy, trust, and generally human welfare have been raised. Designing for moral values has become increasingly important for development of ICT. In a multitude of systems, e.g., electronic health records or social networks, human values play a role and are sometimes violated. One difficulty is that the (long-term) effects of a technology and the impact on people's values can only be assessed fully after the technology has been developed. Hence, the technology is already in use, but policies are not in place to avoid harm. Therefore, including considerations of human values and systemic effects of technology early on in the design process is imperative.

In the area of ICT, software engineers are among others responsible for ethical development of ICT systems. However, many engineers are not aware of how their design decisions influence the ethical impact of technologies and even the ones who are aware have seldom received training on how to account for values in design. One may claim that there are other professionals to account for human needs in ICT design, such as human factors and user experience engineers. The first are concerned with aspects of the user-system interface in terms of matching the human perception and physical abilities, while the latter consider a broader set of

needs and effects of different designs on the experiences of users. Yet, sometimes even the decisions made on an algorithmic level, such as a single number setting a threshold in a computer program, can impact human lives as demonstrated in the following case.

Consider a software producing MR scans used by doctors to estimate blood volume of the heart in order to diagnose heart disease (Kraemer et al. 2011). The software image depicts the difference between blood and heart muscle tissue with light and dark gray colors. However, the borders are not sharp. Therefore, a segmentation algorithm is used that sets a numerical threshold (all gray values above the threshold are "light" and all below are "dark") to distinguish blood from muscle tissue. Due to the noise in the MR scan, there is no correct value for this threshold. The choice of the threshold influences the estimated blood volume and, eventually, the diagnosis in borderline cases.

As Kraemer and colleagues explain, it is practically impossible to design algorithms with a 100 % success rate. Existing algorithms may produce false negatives and false positives. In result, either not treating an ill patient or administering a treatment that is not necessary can have devastating consequences for patients. To decide on the threshold, and thereby favoring either false positives or false negatives in borderline cases, requires an ethical consideration (e.g., regarding patient safety). Depending on the circumstances, e.g., the severity of the disease or the effect of treatment, either case may be preferred in one context or the other. Such a preference, as argued by Kraemer and colleagues, should be in the hands of the doctors, i.e., the users of the system, and not the engineers.

The case highlights the importance of ethical considerations in ICT design, which has been and continues to be the focus of diverse disciplines, including philosophy, social sciences, design, and human-computer interaction. Within the different fields, the focus ranges from analytical and critical studies of the impact of ICT on people's lives to proactively developing and applying methods to account for values in ICT design. The goal of this chapter is to provide an overview of these approaches and methods as well as give examples from practice and discuss open issues. Before getting there, we will start with a brief introduction to ICT, its history, and the central values implicated in ICT design.

## ICTs

ICT has been, since the 1980s, and still is a major growing sector in the EU. "[T]he ICT sector represents about 5 % of the EU economy, but it generates 25 % of total business expenditure in Research and Development (R&D), and investments in this sector account for 50 % of all European productivity growth" (Sallai 2012). As we will see in the following, ICT is a broad term comprising a diverse set of technologies, applications, and services that is still rapidly growing due to new developments in areas such as digital technology, sensor technology, and LED technology.

## Definition and Examples

The term "information and communication technology" (ICT) is an extension of "information technology" (IT), i.e., "[t]he study or use of systems (especially computers and telecommunications) for storing, retrieving, and sending information" (Oxford Dictionary 2014). Data is stored today either magnetically (e.g., on hard discs) or optically (e.g., CD-ROMs). It is organized in large databases managed by database management systems that support retrieval of data, e.g., through the structured query language (SQL). Sending of information can be done in the form of broadcasting (unidirectional) or telecommunications (bidirectional). Data is transmitted in computer networks using protocols, such as SOAP for Web services, and languages such as Extensible Markup Language (XML).

While both terms, IT and ICT, are often being used as synonyms, ICT stresses the role of unified communications. The digital convergence based on the large-scale development of digital technology was triggered in the early 1980s, when telephone networks began to be digitalized, and information technologies were used in digital terminal devices as well as public transmission networks. The two originally distinct sectors of information technology (mainly concerned with mainframes back in those days) and telecommunications (mostly concerned with telephony) were thereby merged. Other additions were the electronic media and content-producing sectors (see Sallai (2012) for a detailed description of this convergence process). Due to the widespread use and development of the Internet as well as IP-based services, ICT is today more in focus than ever.

ICT has become an umbrella term for a diverse set of technologies, including radio, television, mobile devices, hardware and software for computers and networks, satellite systems, as well as many services and applications that are connected to these. Examples of ICT that are in widespread use today are archiving and documentation systems, knowledge-based systems (e.g., classification systems or decision support systems), online communication systems, Web services, and mobile phone applications – to name the important ones. As diverse as the technologies are the application domains, such as social networking, education, healthcare, collaboration at work, personal communication, eCommerce, sustainability, and many more.

With such a large scope and continuous rapid technological developments, the field is expanding and changing as we write, and new ICTs are emerging posing special hurdles to ethical investigations of their impact, an issue we will return to at the end of this chapter.

## Brief History

There are many ways to describe the historical development of ICTs. One could, for instance (similar to the introduction above), describe the type of computational artifacts and their characteristics having developed from large mainframe computers used by several people over personal computers used by a single user to

computing being embedded into the environment and therefore becoming less visible to its users. Instead, we suggest a different perspective on the history that is more relevant to this book, namely, a perspective focusing on ICT development and its openness to its environment, i.e., to users, to social contexts, and to society at large.

In the early days of information technology development from the 1960s to the end of the 1970s, the prevailing effort was in developing innovative technologies. We could observe a technology push, where single users and their contexts were of little importance. Computers were not mass products at the time, but only used in companies to support work tasks by expert users with special training. This situation changed with the first personal computers entering the market in the 1980s. Gradually, computers turned from mainly being problem-solving machines in the work sphere into private devices for people to execute tasks at work and later even at home. These technically untrained people ran into many difficulties, and it soon became obvious that computer interfaces had "usability" issues – the term "usability" just arising. This second stage of ICT development was based on realizations that computers were operated by human beings and in real-world organizations and consequently led to a turn towards the user. Approaches like user-centered design (Norman and Draper 1986) were established to ensure human requirements and needs to be met in the design of the new generations of computational artifacts. Throughout the following decades, we saw the standardization of usability as well as user-centered design and with the rise of the Internet and mobile devices a move towards user experience, the latter incorporating more than usefulness and efficiency, but also joyful aspects of using computers for many different purposes.

In the third stage, and that is where we are now, we are dealing not only with user-computer dyads but also with large networked systems and computation embedded into the environment. These developments gave rise to larger societal questions. A turn from user needs to citizen values can be observed. Societal and individual values take an active role driving innovative ICT design today.

## Central Values

The focus on particular values in ICT design is linked to the historic development outlined above and to the focus of the approaches to design for values outlined in the following section (e.g., democratic values were in the focus of participatory design). While there is no limitation to the set of values relevant to ICT, some researchers have provided lists of values that have a "distinctive claim on resources in the design process" (Friedman and Kahn 2003) or at least can be seen as heuristics for ICT designers (Friedman et al. 2013b). These lists included human welfare, ownership and property, privacy, freedom from bias, universal usability, trust, autonomy, informed consent, accountability, identity, calmness, and environmental sustainability. Whether values should be singled out as particularly worthy for consideration (Borning and Muller 2012) or whether

such lists may bias system designers towards these values (Le Dantec et al. 2009) is in focus of ongoing debates. In this chapter we do not intend to take a stance on these issues.

As we do not have the space to provide a comprehensive list of values and descriptions, we describe those values that have often been in the focus of ICT design until now. That said, one should keep in mind that the central values are linked to the developments of a certain time and in a certain cultural context. With new ICTs and the widespread use in different cultures, other values may surface and be of central relevance.

## Privacy

Privacy – a basic human right (Movius and Krup 2009) – and, in particular, its link to data protection, has been a central concern in ICT research as well as in public debates (e.g., about government surveillance of domestic communications). Due to increased computing power, data storage, and ubiquitous sensing of personal data, accumulating and linking personal information has become easier. Despite the benefits (e.g., personalization), this also leads to compromising individual privacy. "Privacy can be endangered by camera surveillance, monitoring of internet communications, . . . availability of individual medical information in the public health system,. . . [etc.]"(van den Hoven et al. 2012).

Since privacy and designing for privacy is discussed in detail in one of the chapters in this book (see chapter "► Design for the Value of Privacy"), we will not go into detail, but highlight the central aspects Warnier, Dechesne, and Brazier discuss in their chapter. The concept of privacy entails three different aspects: (1) being left alone and free from intrusion, (2) being able to control information about oneself, and (3) not being tracked, followed, or watched in private space. These conceptions are, however, not absolute, but relative to a person's circumstances. Privacy and its perception may, for instance, be different in public spaces and one's home. However, this is exactly where the lines blur when it comes to online communications. Are blogs, websites, or social networking sites considered public realm by their users? Often people are less concerned with their privacy when online, but more with the possible harms and discriminations resulting from a misuse of their data. To protect from such misuse, EU directives (e.g., EU Directive 95/46/EC) have been put into place.

Regulations and laws try to protect people through rules enforcing "(i) transparency (How is data stored/processed?), (ii) purpose (Why is data stored/processed?), (iii) proportionality (Is this necessary for this goal?), (iv) access (What do they know about me, can I change it?), and (v) transfer (Who else has access?)" (► Design for the Value of Privacy, in this book). Besides regulations and laws, design of ICT can take a proactive stance in ensuring data protection and privacy of users, e.g., by taking into consideration informed consent in the design (see, e.g., Friedman et al. 2013), thereby being transparent about what information is stored and giving users the opportunity to opt out or adjust storage settings.

## Security

ICT security has long been the interest of large companies storing sensitive data. With the widespread introduction of ICT into home environments, where sensors collect data, consumer devices are linked and connected continuously to the Internet, people store data in the cloud, and many users are not aware of the security of their systems, we are in need of new strategies to ensure people are not harmed by security attacks. Denning and colleagues stated that "[d]evices in the home will likely incorporate varying degrees of security defenses, due in part to oversights by designers and developers, but also due to the costs associated with implementing security measures" (Denning et al. 2013). They provided a framework that systematically identifies key security risks within the home computing context including ICTs' exposure to attacks and their attractiveness, possible impact on human assets, and important security goals to address.

## Ownership

"[O]wnership can be understood as the general right to property, which, in turn, entails . . . the right to possess an object, use it, manage it, derive income from it, and bequeath it" (Friedman and Kahn 2003). This definition is straightforward in the realm of physical objects. However, when it comes to ICT, many questions about the ownership of digital data such as comments in online forums, cookies in browsers, or user profiles are raised. It is generally known that Web companies collect data about their users, for instance, for personalization purposes. However, it is often unclear to users what type of data is stored about them, what is done with it, and whom it is being shared with. Even more complicated is ownership of social media content, which may be created collaboratively and thereby jointly owned. What rights with regard to saving, sharing, publishing, or removing apply for collective data? People's opinions on these matters differ, as Marshall and Shipman (2011) showed. ICT developers have a strong influence on the execution of users' ownership rights through the designed functionalities (which in turn may influence people's attitudes towards ownership in the long term).

## Universal Usability

Usability, i.e., the efficient, effective, and satisfactory use of ICT systems, has been a key concern in ICT design since the 1980s. Universal usability refers to the idea that every person should be able to use information systems successfully, which entails (1) access to hardware, software, and networking; (2) accommodating users' backgrounds (e.g., income, abilities, knowledge); and (3) "bridging the gap between what users know and what they need to know" (Friedman and Kahn 2003). In particular, the second aspect is being addressed in a design approach

called Design for All (or Universal Design) to ensure a system allows for human diversity, social inclusion, and equality (EIDD Stockholm 2004).

## Autonomy

Autonomy in ICT refers to users' control over the technology, in particular "the right things at the right time" (Friedman and Kahn 2003), in order to plan and execute their actions in a way that achieves their goals. Four aspects in ICT design can support or hinder user autonomy (Friedman and Nissenbaum 1997): (1) system capability (e.g., offering functions to opt out), (2) system complexity, (3) misrepresentation of the system, and (4) system fluidity (i.e., being adaptable to the user's goals). Especially with the recent advance of ambient intelligent systems, which sense the user and context and sometimes act on behalf of the user (e.g., within telecare a system may send data about the user's state to a mobile phone of a relative), user autonomy is a central value at risk and, therefore, needs to be considered carefully in the design.

## Trust

When analyzing trust in ICT, an important distinction is to be drawn between objects of trust. Are we – as users – to trust an ICT system, its designers, the company behind an ICT service, or another person we are communicating with through an ICT system? In early research on "cyberspace" ('Schneider 1999), trust and trustworthiness were concepts attributed to a system's functioning in terms of correctness, security, reliability, safety, and survivability. These are surely important aspects of systems, especially those providing information in high-risk, safety-critical domains. However, this limited view neglects fundamental characteristics of trust as a social value, i.e., experiencing good will (between people), extending good will to others, being vulnerable, and experiencing betrayal (Friedman and Kahn 2003). In examining trust online, Friedman, Khan, and Howe (2000) distinguished two contexts for trust online: eCommerce, where it is often hard for users to judge a company's good will and the harms associated with a transaction, and interpersonal relationships, for instance, in online social media, in which violations of trust may cause psychological harm.

## Accountability and Responsibility

Information systems, especially in high-risk domains (e.g., air traffic, medical decision making, military applications), can put human lives and well-being at risk in cases of failure. The question is "who is to blame in these cases?" Studies have shown that people attribute agency and responsibility to computers (e.g., Friedman 1995) despite the fact that they cannot be liable. This could, among

other things, either be due to the complexity of the system preventing users from understanding the impact of their actions or due to systems mimicking human agency (Friedman and Kahn 2003) (especially in agent technologies following the believes, desires, and intentions paradigm, which sometimes even have interfaces with human features (e.g., avatars)). It is up to the designers of ICT to mitigate these issues, e.g., by making the decision-making process of systems more transparent to allow users to estimate the impact of their actions based on recommendations from the system.

New ICTs allowing users to have knowledge that they could not access before can also create new responsibilities for users. Consider the case of ambient assisted living systems sending information and warnings to informal caregivers about the health and safety status of a senior resident (Detweiler et al. 2012b). In this scenario caregivers become responsible for taking action when health threats are detected.

### Human Welfare

Friedman and Kahn (2003) have distinguished between three types of human welfare: physical welfare, material welfare, and psychological welfare. Safety-critical systems can potentially harm the physical welfare of people, i.e., bringing injury or in the worst case death, when the hardware or software fails or wrong design decisions are made (see example of medical image technology given in the introduction). Material welfare refers to physical objects and human economic interest, which is at stake, if personal data (e.g., financial) gets damaged or stolen. Last, psychological welfare (the mental well-being of a person) can be promoted or harmed through ICT, in particular, in the social Web, where friendships and communities can be formed, but also risks of cyber bullying are present.

## Approaches to Design for Values in ICT

### Overview of Approaches from Philosophy, Social Sciences, and HCI

Several important approaches to dealing with human values and ethics in ICT have emerged in different disciplines. This section intends to give an overview of the range of approaches. It needs to be emphasized, however, that the approaches vary strongly in terms of the amount of research done to develop the approach itself as well as their impact on research and practice. While some comprise whole research fields with vast amounts of published research (e.g., CSCW or participatory design), others (e.g., values at play, worth-centered design) are frameworks or methods that have been suggested in the literature and used in few projects.

#### Computer Ethics

One approach is computer ethics (Johnson 1985), part of the practical philosophy tradition, which tries to understand the impact of computing technologies on

social life. To that end it aims, on the one hand, to utilize existing moral theories (see, e.g., the conceptualization of trust online in Friedman and Kahn (2003)) and, on the other hand, to extend the boundaries of existing ethical concepts based on new ICT development (e.g., concepts on online privacy in social networks). While computational ethics can provide useful insights to our understanding of ICT and human values, it is not a design approach (Friedman and Kahn 2003) and, therefore, does not offer guidance for developers for the technical implementations of ICTs and design trade-offs implicating more than one value.

### Social Informatics

Social informatics is "the interdisciplinary study of design, uses, and consequences of information technologies that takes into account their interaction with institutional and cultural contexts" (Friedman and Kahn 2003). A key theme within social informatics research is the "social context" of ICT. Social context is believed to have a large influence on the way people will use (or not use) ICTs and in turn also influence work, organizations, and social structures. "Social context does not refer to some abstracted "cloud" that hovers above people and information technology; it refers to a specific matrix of social relationships. [In the example of Lotus Notes], social context is characterized by particular incentive systems for using, organizing, and sharing information at work. . . . [D]ifferent groups . . . have different incentives to share information about the project know-how, and, thus, how they use or avoid Lotus Notes" (Kling 1999). Besides the adoption and diverse use of ICT, social informatics researchers focused on the consequences of computerized information systems and especially their infrastructures on the organization of work, as, e.g., shown in the case of classification system (Bowker and Star 1999).

In social informatics, computerized information systems are seen as socio-technical systems that lend themselves to analyses that go beyond the technical artifacts. Through analysis of the social context by using workplace ethnography (Simonsen 2009) or participatory design methods (see below), the design of ICT can be informed in ways that account for the social context and values of the future users.

### Computer-Supported Collaborative Work

A field that originally focused on ICT in the workplace is computer-supported collaborative work (CSCW). It is an applied field and by that has a stronger focus on the design of new ICTs to support group work than the approaches above. Central values in CSCW are closely linked to group activities, e.g., cooperation as the overarching value (Friedman and Kahn 2003), but also related values such as privacy, autonomy, ownership, commitment, security, or trust. To understand the social settings and values at play, CSCW researchers have commonly made use of ethnomethodology (see, e.g., Crabtree 2003). In the last decade the field has expanded to more intimate groups, communities, and societies in non-work settings. With the move into domains of intimate relationships between people (e.g., within families or romantic relationships), researchers have come to notice the increased ethical implications. "The value-laden and emotional nature of these new

explorations complicate the moral landscape and require us to revisit the ethical aspects of questions like "how should we study?" and "what should we design?"" (Branham et al. 2014).

### Participatory Design and Co-design

A design tradition that has engaged with human values is participatory design (PD). Dating back to the early 1970s, when labor unions had a strong influence in Scandinavia and helped to put into place a co-deterministic agreement empowering workers to influence decisions of technology introduction into the workplace (Friedman et al. 2003), PD has values of workplace democracy, quality of work life, and human welfare embedded into it. Within the tradition of PD, methods to allow for the active engagement of workers into the design of ICT systems have been developed, among others, contextual inquiry (Holtzblatt and Jones 1993), cooperative prototyping (Bødker and Grønbæk 1991), and future workshops (Kensing and Madsen 1992).

Closely linked to the sociopolitical setting in which PD arose, methods focus on envisioning futures involving changes in the social, technological, and political environment. Over the years approaches to involve users actively in the design of ICT have prevailed in human-computer interaction and software engineering according to the belief that "active user involvement in the software development process leads to more useful and usable software products" (O'Neill 2000). Co-design (Sanders and Westerlund 2011), a more recent participatory approach, focuses on services and products in diverse settings and not only the workplace. While democratic values are still embedded in this approach, it focuses more strongly on the role of creativity in design processes.

### Values in Design

Another strand of research into values embodied in technologies is values in design (Nissenbaum 1998, 2005). Triggered by investigations of bias (Friedman and Nissenbaum 1996) and user autonomy (Friedman and Nissenbaum 1997) in computer systems, Nissenbaum set out to develop a methodology for system engineers to embed values into software. Based on the metaphor of juggling balls, the methodology suggests the system engineer keeps in play three distinct modes of knowledge: the technical mode (i.e., scientific and technical know-how on design specifications to realize given values), the philosophical mode (i.e., characterizations of values and rationale for commitments to particular values), and the empirical mode (i.e., investigation of whether attempt to embody values was successful). Results from all three modes are meant to be integrated.

### Values at Play

Values at play "is a systematic methodology for discovery, analysis, and integration of values in technology design" (Flanagan et al. 2005) that takes a game design approach, meaning that it is investigated how designers can create and use games to explore, question, or affirm cultural and moral values. The methodology has three major stages, i.e., discovering values relevant to the project, deciding which to

integrate into the design, then translating them into design features (through iterative cycles of prototyping), and, last, verifying that the values are embedded in the project. To make game designers more sensitive towards the values at play in games, typical exercises would be to use value cards (Flanagan et al. 2007) to spur discussions about how specific values are promoted or violated in existing games.

### Value-Sensitive Design

Value-sensitive design (VSD) is a design framework aiming to incorporate knowledge of the ethical impact of a technology into a design process. VSD "is a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" (Friedman et al. 2013). To that end, it provides an iterative three-part methodology consisting of conceptual investigations, empirical investigations, and technical investigations with direct and indirect stakeholders, their values, and technical developments at the core of these investigations. As there is a complete chapter devoted to VSD in this handbook (see chapter "► Value Sensitive Design: Applications, Adaptations, and Critiques"), we do not go into depth here. That said, we would like to emphasize that VSD has been a very active research area in the past 20 years and is still a growing area continuously developing new methods to put value-sensitive design into practice (see section "Methods to Account for Values in ICT Design" for VSD methods).

### Privacy by Design

Recent rapid and complex technological developments, especially with regard to the analysis of big data, and associated privacy concerns provided the motivation to embed privacy into the architecture and design of new technologies. The resulting Privacy by Design framework is an exemplification of the VSD approach with the goal to " inscribe privacy protection into the analytical technology by design and construction, so that the analysis takes privacy requirements in consideration from the start" (van den Hoven et al. 2012). Privacy by Design has seven foundational principles (Cavoukian 2009): (1) being proactive, (2) privacy as default, (3) privacy embedded in design, (4) commitment to functionality in a positive-sum strategy (e.g., avoid privacy vs. security trade-offs), (5) life cycle management of data, (6) visibility and transparency to users and providers, and (7) respect for user privacy.

### Worth-Centered Design

Worth-centered design is a "development framework that supports a focus on value from the initial identification of product opportunities to the installation and operation of digital products and services" (Cockton 2006). Once called value-centered design, it originated from the context of usability and, in particular, evaluation of usability with respect to contextual fit (something occurring as a usability problem in one usage context may not be a problem in another). "Unlike value-sensitive design … VCD has no roots in moral considerations, although it cannot avoid them" (ibid). To avoid confusions between the two frameworks,

value-centered design was renamed and developed into worth-centered design, as it focuses on the *worthwhile* irrespective of ethics.

## Values in Existing Software Development Models

The design and development of ICT systems encompass a wide range of disciplines including the ones already named. Another core discipline is software engineering, which deals with the design, development, and maintenance of software. Several software development models have been developed over the years, ranging from early waterfall models over iterative or cyclic models to the more recent agile development models. Whether iterative or not, most models share a set of core activities, i.e., requirements elicitation, specification, architecture creation, implementation, testing, deployment, and maintenance. In particular the first phase, where requirements of a new system are defined, seems suited for analyzing different stakeholders and their values. Requirements engineering attends to "soft issues" such as politics and people's values, although dealing with soft issues is problematic as there is little guidance on how to do so (Thew and Sutcliffe 2008). Therefore, Thew and Sutcliffe provided a taxonomy of users' values, motivations, and emotions to help elicit and analyze these issues in requirements engineering and give guidance in how to use the taxonomy in conjunction with interviews and ethnographic methods. Some support on how to adjust the design process to elicited values is given, e.g., "organizing the team composition in response to aesthetic needs (e.g. include aesthetically aware designers)" (Thew and Sutcliffe 2008). However, there is a lack of specific methods supporting the translation of values into concrete designs or dealing with value trade-offs.

Agent-oriented models commonly use the concept of goals. In TROPOS, for instance, stakeholders and their goals are identified in the early requirements phase. Morandini et al. (2008) suggest beginning this phase by asking questions such as "who are the stakeholders in this domain? . . . what are their goals and how are they related to each other?" ScenIC (Potts 1999) refers to two types of goals: objectives, expressed as a "trajectory of improvement," and tasks that are "stated in terms of achievement of a state or performance of an action." Potts suggests obtaining the system's goals from mission statements, questions to stakeholders, and the like. He further provides a lexicon of verbs useful for identifying objectives and tasks. For example, an objective of goal achievement might be improving a condition, with which the verbs "improve," "reduce," and "maximize," among others, are associated. In both TROPOS and ScenIC, the concepts of goals are similar to the concept of values. However, "[v]alues are not goals . . . A value is a judgment, though very general and vague. It says of something that it is good or bad. A goal is a regulatory state in someone's mind" (Miceli and Castelfranchi 1989, p. 179). The authors illustrate a further important feature of values in discussing the difference between values and norms: "Values in fact offer grounds for, or give rise to norms. Hence the "normative" facet of values: If something is good, it should be pursued" (Miceli and Castelfranchi 1989, p. 181). If we represent values as soft goals, the evaluative

aspect ("X is good") and the normative aspect ("X should be pursued") are lost. Represented as a soft goal, a value becomes something that can be satisficed (i.e., sufficiently satisfied). Not achieving a goal is not morally wrong as such. Violating a value, on the other hand, can be seen as morally wrong. Not taking these aspects into account could lead to problems once the design has been implemented and put into practice.

A current trend in software development is the use of agile development methods, such as SCRUM and extreme programming (XP). These commonly represent the features to be developed to create business value as user stories in the form "As a < role > I would like to < goal/desire > so that < benefit>." Recently, first attempts have been made to inject value consideration into user stories by extending the template to "As a < stakeholder > I want < stakeholder need > so that my < value > is promoted/supported when < concrete situation>" (Detweiler et al. 2014).

## Methods to Account for Values in ICT Design

As an applied field, ICT design makes use of methods that can inform the design processes in terms of surfacing user needs and values, (technical) constraints, and supporting creativity and imagining the future. Depending on the respective discipline, methods can be formal (as in requirements engineering) and software development (e.g., value user stories in the previous section) or rather open-ended to inspire creativity in design. To account for values in ICT design, a range of methods is used with different purposes, e.g., identifying values relevant in the design context, creating working definitions, analyzing and resolving value trade-offs, envisioning future designs and their impact on values, transforming value analyses into concrete requirements, and ultimately working designs.

VSD, in particular, entails three types of investigations, i.e., conceptual, empirical, and technical, which – when integrated and used iteratively – give a holistic view on stakeholders, values, and technical issues. Many existing methods from the social sciences have been used in these investigations and new methods have been developed to answer specific questions in VSD projects. In conceptual investigation the focus is on questions about the stakeholders, their values, and value trade-offs. A method to identify stakeholders could be a document analysis of design briefs or grant proposals. The latter are also often resources of value claims. Besides identifying stakeholders' values, an important task in this step is to create working conceptualizations of specific values. A method to do so could be conducting a systematic review of philosophical literature. In case technologies already exist in the design domain, technical investigations can be added to analyze existing artifacts' value suitabilities, i.e., whether they support or hinder the identified values.

Conceptual investigations and these types of technical investigations are inherently limited, as they do not involve direct observations of the design context. For this we need to conduct empirical investigations. VSD does not prescribe concrete

methods for empirical investigations, but states: "the entire range of quantitative and qualitative methods used in social science research is potentially applicable here, including observations, interviews, surveys, experimental manipulations, collection of relevant documents, and measurements of user behavior and human physiology" (Friedman et al. 2013). We will not go into detail of these methods, as they are well established and described in social science literature. Instead, we will focus below on specific methods that VSD researchers have developed to better suit VSD investigations. The methods we describe in the following are used to answer different types of questions arising in the three types of investigations including, for instance, the identification of relevant stakeholders, value elicitation, or design methods.

### Stakeholder Analysis

A method usually employed in the beginning of a design project is *stakeholder analysis*. While many approaches focus merely on users of ICT, VSD takes into account two classes of stakeholders: direct and indirect ones. "Direct stakeholders refer to parties – individuals or organizations – who interact directly with the computer system or its output. Indirect stakeholders refer to all other parties who are affected by the use of the system" (Friedman et al. 2013). For instance, in a medical record system, direct stakeholders could be doctors, hospitals, nurses, and insurances. An indirect stakeholder group would be patients (who do not directly interact with the computer system, but whose privacy could be compromised by it).

In a stakeholder analysis, a design team answers the following questions: Who are the direct and indirect stakeholders affected by the design? What values are implicated for each group? For both questions, the team creates lists that can be referred to throughout the design process. Importantly, direct and indirect stakeholders refer to roles. One person could take on different roles. While a doctor using an electronic health record system is a direct stakeholder the majority of the time, he or she could also be an indirect stakeholder, when being ill and taking on the role of a patient.

### Value Elicitation

Results from a stakeholder analysis are usually elaborated through empirical investigations of the stakeholders' needs and values. Identifying or eliciting stakeholders' values is, however, rather tedious, as it is difficult for people to express their values. Within the social sciences, value questionnaires and inventories have been around for some decades, most notably the Rokeach Value Survey (RVS) (Rokeach 1973), Schwartz Value Survey (SVS), and the Portrait Value Questionnaire (PVQ) (Schwartz and Bilsky 1990). However, when eliciting values for a specific design case, such questionnaires, which give abstract value priorities of people, are not sufficient to explore the complex value systems, dependencies, and in situ values relating to the design in question. Several other social science methods have therefore been adapted and employed by VSD researchers to study values in situ: interviews in situ (Friedman et al. 2006, 2008a, b), surveys in situ (Friedman et al. 2008a, b), physiological measurements in situ (Kahn et al. 2008),

and diaries (Friedman et al. 2008a, b). These studies showed the worth of value deliberations relative to the use context of a certain technology (Friedman et al. 2006). Recently, the development of more specific value elicitation methods has been the focus of researchers (e.g., Detweiler et al. 2012a; Pommeranz et al. 2011). These include direct approaches such as specific digital tools to support stakeholders to reflect on their values related to a certain context in everyday life (Huldtgren et al. 2013) or photo elicitation interviews (Le Dantec et al. 2009) and indirect approaches such as the approximation of users' values through preferences on tasks (e.g., in the work context of nurses (Koch et al. 2013)) or content analysis of online data (e.g., from tweets (Koepfler and Fleischmann 2011)).

### Value Sketches

Sketching is often used in design work to uncover knowledge for "physical and conceptual structure" (Woelfer et al. 2011). Value sketches in particular are meant to emphasize participants' values. For instance, in a study with homeless youth, Woelfer et al. (2011) asked participants to sketch out their perception of safety in different parts of the city by day and by night. To that end they were given two identical maps of the city and were asked to use red and green (graphic and textual) marks for safe and unsafe areas, spots, and paths. Through detailed coding and analysis of the sketches, Woelfer and colleagues could retrieve a detailed picture of temporal and location-sensitive perceptions of place, mobility, and safety for each stakeholder group.

### Value Dams and Flows

After eliciting detailed accounts of values, it is analyzed in what way ICT can support or hinder the values of particular stakeholders. For this, the "value dams and flows" method is a useful tool. Value dams are "technical features or organizational policies that are strongly opposed by even a small set of stakeholders" (Miller et al. 2007). Hence, implementing ICT (or policies) that incorporates such dams will hinder the adoption of the ICT. Therefore, such features should be avoided – also to protect the welfare of minorities. Value flows, on the contrary, are features that many stakeholders support. Even if not needed for the system to function, such features will increase the adoption and perceived value of the system. "Value dams and flows" can also be used to understand stakeholders' value tensions.

### Value Scenarios

Value scenarios (Nathan et al. 2007) combine the narrative power of traditional scenarios (Rosson and Carrol 2002) with five new key elements that help to engage in (ethical) issues of long-term and emergent use of ICT: indirect stakeholders (additionally to direct ones), pervasiveness (effects from the widespread adoption of the technology), time (long-term effects), systemic effects, and value implications. By describing possible positive and negative effects and value tensions that come along with widespread adoption, value scenarios support technologists and policy makers to consider the creation and introduction of new technologies.


**Fig. 1** Envisioning Cards (*front side* on the *left* and *back side* on the *right*). (Source: VSD lab, University of Washington (UW), permission to reprint the image and copyright remains with UW. See also: www.envisioningcards.com)

## Envisioning Cards and Security Cards

To put technology development in a broader socio-technical and long-term perspective, by highlighting "diversity, complexity and subtlety of human affairs, as well as the interconnections among people and technologies" (Friedman and Hendry 2012), the Envisioning Cards toolkit (Friedman et al. 2011) provides a promising means.

Envisioning Cards incorporate similar elements to the value scenarios: stakeholders, time, values, and pervasiveness (see Fig. 1). The Envisioning Cards set is a versatile tool that can be used in many design processes including ideation, co-design, heuristic evaluation, and critique or as an educational tool. The cards are self-explanatory and open for different types of use, which makes them equally accessible to designers, technologists, and end users and supports them in ethical reflection.

Similar, to the Envisioning Cards deck, a Security Cards deck has recently been developed (Friedman et al. 2013a) as a brainstorming toolkit specifically for considering computer security threats. The cards can be used in different contexts, as an educational tool for students, as a design tool for software developers, or as a communication tool between developers and management.

## Value-Sensitive Action-Reflection Model

In a design case described in more detail in the following section "Safety and Homeless Youth," we have sought to provide stakeholders with tools to articulate their needs, values, and visions of yet-to-be-built tools (mobile ICT to promote the value of safety for homeless young people). "On reflection, [the researchers] recognized that the method [they] developed for one particular study could be generalized and [they] have sought to do so through the Value Sensitive Action-Reflection Model" (Yoo et al. 2013). The model (see Fig. 2) was used in co-design sessions with different stakeholders to evolve the co-design space. It can do so in several ways. By using designer prompts (design tools such as the Envisioning Cards) and stakeholder prompts (tools generated by other stakeholders, e.g., value

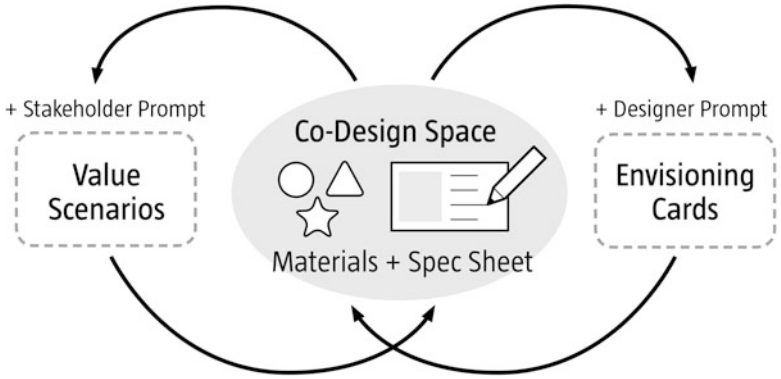



**Fig. 2** Value-sensitive action-reflection model (Yoo et al. 2013)

scenarios), the model injects and facilitates moments of reflection on the co-design activities (e.g., cooperative prototyping). By using two different types of prompts, the model leads to divergent thinking.

### Value Levers

The methods discussed up to this point focused mostly on identifying values and understanding value complexities through engaging directly with stakeholders. Another aspect of designing ICT with values in mind is the value consideration within design teams. To that end Shilton (2012) has investigated social values and design practices within a research lab and introduced "*value levers*: practices that pried open discussions about values in design and helped the team build consensus around social values as design criteria" (ibid, p. 376). The conclusion of her long-term ethnographic field work points to the importance of the structure of design labs and the forces of laboratory practice influencing the creation of value levers, such as "working on interdisciplinary teams, experiencing internal prototype testing, designing around constraints, advocacy by leaders and a values worker, navigating institutional mandates, and gaining funding, [which] promoted social values in design" (ibid, p. 393).

## Experiences and Examples

### Values in the Filter Bubble

A popular example of values being affected by ICT is the case of online information filtering leading to a filter bubble – a term coined by Pariser (2011). The Web 2.0 and information streams from social networking sites such as Twitter, Facebook, and others provide nonstop information from many sources easily leading to an information overload for the users. To manage the information sites employ news recommendation algorithms that filter, select, and provide only that information which is assumed to be interesting or most relevant to the user. Different strategies

exist to do so, for instance, content-based methods, which filter new information based on people's previous choices or collaborative filtering, which takes into account choices of other users with similar tastes. In order to correlate old and new information, systems create detailed user profiles over time containing demographic information, preferences, consumed items, and user context. While it is easy to argue that this data collection and profiling has an effect on people's privacy, that is not the only value at stake.

Pariser (ibid) criticized that overspecialized recommenders, which continuously put information into the foreground that is based on what we are interested in already or what like-minded people share or like, will put users in a filter bubble where they are recurrently exposed to similar views. While the benefits of personalization are apparent, selective information consumption has important negative consequences. Bozdag and van de Poel (2013) recently discussed the impact of online news recommendations on information diversity. They stated that "[p]ersonalization systems, ideally, help users to make choices after carefully weighing all the media content thrown at them and consume only the relevant ones. However, providing people with only agreeable news items may have negative social consequences. Group deliberation among like-minded people can create polarization; individuals may lead each other in the direction of error and falsehood, simply because of the limited argument pool and the operation of social influences." In their following analysis they applied value-sensitive design to conceptually analyze the value of information diversity and provide a technical analysis leading to a number of guidelines for supporting diversity in design of news recommenders. In their conceptual investigation, they used theories from philosophy and media studies to identify three dimensions of diversity, i.e., content, source, and exposure (Napoli 1999). While it is often believed that source and exposure diversity (the number of available and used sources to attain news) lead to greater content diversity (diversity in format, demographic, and idea-viewpoint), this is not empirically proven (ibid). Bozdag and van de Poel (2013) continue to describe a normative framework of diversity that posits that media diversity needs to be externally gauged and that a good media policy has to balance reflective diversity (i.e., "reflecting the distribution of preferences, opinions, allegiances" of the population) and equal access diversity (i.e., "media provides perfect equal attention to all identifiable preferences, streams, or groups, or positions in society"). In order to design for diversity, the conceptualization of the value needs to be translated first into general norms and the norms into design requirements. One such example is reflected in the following quote:

> For instance, reflective diversity can be translated as "user should not get news items that he only agrees with, challenging items must be presented by the algorithm". To use this norm in a Twitter based news recommendation system, news sources can be mapped into different ends of the political spectrum... Later, the design requirement can be specified as follows: maximize the chance of being exposed to different viewpoints on Twitter (specification of the goal) by analyzing type of source the user is subscribed to (specification of means), determining user's political bias with comparing the type of messages one receives and type of messages one sends (specification of means) and then showing him challenging items (specification of means). (Bozdag and van de Poel 2013, p. 1105)

## Care-Related Technology

An area where the introduction of ICT is currently taking place is the care domain. Mainly triggered by the ongoing demographic changes, technology is seen as a key enabler for maintaining high care standards while the number of seniors requiring support is growing and the number of skilled workforce in this sector is decreasing. Unfortunately, a majority of the development efforts represent a technology push and thus do not lead to widespread acceptance of the technologies. Yet, ICT can provide opportunities to promote, for instance, independent living. Telecare is the key technology we are referring to here and the projects in this field are vast (van den Berg et al. 2012). In this sensitive design space with many stakeholders, e.g., seniors with physical or cognitive problems, informal caregivers (partners or family members), professional caregivers, insurances, welfare organizations, and more, different values are at stake and in tension. For instance, the value of efficiency of care may conflict with the well-being and social needs of the patients; the value of safety promoted by telecare may conflict with the value of privacy. Other central values in this field are responsibility, independence, identity, community, quality of life, and many more. Although it has been shown (Sanders et al. 2012) that a violation of core values (such as identity and self-reliance) can lead to a rejection of the technologies, there are few examples of approaches to the design of systems in this area dealing with the value complexities at hand. In two recent publications (van Wynsberghe 2013; Sharkey and Sharkey 2012), ethical issues in the domain of designing care robots were pointed out. In addition, Detweiler et al. (2012b) have discussed the connection between knowledge and responsibility to show that systems for ambient assisted living can lead to moral responsibilities for different types of users. However, most of such work is limited to conceptual analyses and does not provide design cases and guidance accounting for values and value tensions.

In a recent design case (Fitrianie et al. 2013), we have integrated value considerations in the user-centered design of a smart TV platform with services to support seniors at home. One activity entailed the investigation of interpretations of quality of life (a core value and goal of all telecare applications) from different stakeholders, which surfaced very different conceptions (e.g., being without pain, able to help oneself, enjoying daily activities (from a geriatric doctor's point of view (POV)), social contacts, feeling well and healthy (from seniors' POV), freedom and independence (from a social worker's POV), being mobile (informal caregiver's POV), and being oneself (from dementia care organization leader's POV)). It became clear that quality of life was closely linked to one's current situation, preferences, and abilities and therefore differs from person to person.

In a side track of this project, we also looked at ways to increase the technology acceptance through an information system which we designed to promote seniors' values of self-efficacy and confidence in order to empower them to learn about the

potential benefits of telecare without having it imposed by others or committing to a new system (Huldtgren et al. 2014).

## Safety and Homeless Youth

Safety is one of the basic needs of all people. For homeless young people it can be difficult to secure basic needs, while sometimes even managing physical or mental health issues. They encounter unsafe situations in their struggle to meet their needs, often with civility laws being implicated (Woelfer et al. 2011). Other stakeholders in this case, family, service providers, or police want to know homeless young people to be safe, too, yet sometimes the interpretation from these different perspectives may contradict or be in tension with other values (e.g., independence). For homeless young people, mobile phones have beneficial safety functions ranging from functionality, e.g., calling or texting in unsafe situations, to form factors, e.g., held in particular ways the phone may resemble a gun. However, mobile phones may also create unsafe situations, e.g., if homeless young people trespass at secluded power outlets in order to charge their phones. Thus, the use of mobile technology by homeless young people and its relation to safety is multifaceted and a topic worth investigating.

The design case engaged in design work concerning multiple stakeholder groups with different perspectives and values as well as value tensions within and among individuals and groups. We used Envisioning Cards as an iteration step in a co-design activity (following the value-sensitive action-reflection model (Yoo et al. 2013)) with homeless young people, police, and service providers. After the participants created 3D prototypes to keep homeless youth safe, they were asked to select the Envisioning Cards, consider the theme, and refine their designs if needed. The Envisioning Cards stimulated the creative exploration of the design space. They helped participants "to reframe technical problems, to reconsider technical aspects of their designs, and generally to catalyze their technical imaginations" (Friedman and Hendry 2012). Other verbal and visual methods, e.g., value sketches, provided a rich set of data revealing the nuances of stakeholders' perceptions of safety and its situational nature. Through detailed coding and analysis of the sketches, Woelfer and colleagues could retrieve a detailed picture of temporal and location-sensitive perceptions of place, mobility, and safety for each stakeholder group (Woelfer et al. 2011).

Insights retrieved from the use of these value-sensitive methods and the prototyping activities with stakeholders broadened the design space. Instead of focusing only on technical features and form factors of mobile phones and app development, the process led to considerations of the social context and design socio-technical solutions. One representative example was the idea to provide power supply or backup phone at service providers' station, thereby promoting safety for the homeless young people, who would not have to trespass at secluded

power outlets and at the same time give service providers the opportunity to be in touch in a nonintrusive way.

## Open Issues and Future Work

### Methodological Issues

Approaches such as VSD – although being around for 20 years at the time of writing – are still evolving, and it is important to reflect critically on the premises of the approaches. In the following, we look at three aspects: values, methods, and expertise in the design team.

#### Values

How can one identify the relevant values in a given design case? VSD, for instance, refers to a broad definition of values as "what a person or group of people consider important in life" (Friedman et al. 2013). In result, ICT developers face a big task in identifying the values to focus on in a design case. They also have to be aware whose values these are (see distinction on explicitly supported values, stakeholder values, and designer values in (Borning et al. 2005) and how interpretations of one value may differ between stakeholders). In the VSD literature (as well as in section "Central Values") values with ethical import have been mentioned. Such lists of values have been presented from having "a distinctive claim on resources in the design process" (Friedman and Kahn 2003) to being heuristics for designers. Other researchers, however, have pointed to a danger of a bias towards these values (Le Dantec et al. 2009).

Another issue is the universality of values, which is interesting from a philosophical view. Friedman et al. (2013) stated that VSD "builds from the psychological proposition that certain values are universally held." Another perspective is moral relativism, which states that "the truth or justification of moral judgments is not absolute, but relative to some group of persons" (Gowans 2012). Certainly arguments exist for and against either position, but a full discussion is beyond the scope of this chapter and it is not our intention to argue for one or the other.

Borning and Muller (2012) concluded in a related discussion that the existence of universal values has little impact for the practical application of VSD. They say that as values play out sufficiently different in each design context, universal designs accounting for a certain value are not attainable. Instead, it is important that the values at stake are identified and analyzed carefully as well as defined with respect to the particular context and new design solutions for the given context have to be created. That this is challenging and also calls for reflection on the lenses we use to analyze our empirical data is shown by an interesting cross-cultural design case (technology use in long distance relationships in an Arabic context) provided by Alsheikh et al. (2011). The example shows that using different theoretical lenses (in their case Western, traditional Islamic, feminist Islamic) to analyze ethnographic data would lead to very different design considerations.

### Methods

While people usually agree that accounting for human values in ICT design is good, how to elicit values from stakeholders and how to get from conceptualized values to concrete functionalities in ICT often remain unclear. Originally, "the entire range of quantitative and qualitative methods used in social science research" (Friedman et al. 2013) was suggested as a pool of methods to empirically analyze values. While Borning and Muller (2012) suggested examining the value suitabilities of these methods, others (Le Dantec et al. 2009; Pommeranz et al. 2011) proposed that more specific methods are needed to capture values as lived experiences and to give stakeholders the power to express and share their comprehension of values.

The second methodological hurdle, i.e., getting from values to designs, is one that seems to be underexplored in the literature. The Envisioning Cards and Security Cards provide some guidance for developers in practice; however, concrete support for implementations is lacking. One way to provide more detailed advice to developers could be to analyze existing VSD cases and derive value-sensitive design patterns (similar to the concept of design patterns in software engineering) from successful cases and best practices (Detweiler and Hindriks 2012).

### Expertise in the Design Team

Related to the question of which methods to apply is the issue of what competences should exist within an ICT development team. Looking at the case presented in the introduction on ICT producing images used to make medical diagnosis, what competencies are needed in design teams, which develop such high-stake applications, to understand and analyze the ethical impact a single numerical threshold value may have?

Much research on values in ICT is conducted by HCI researchers or computer scientists, yet it would be beneficial to form interdisciplinary teams to carry out empirical research or conceptual analyses of values, respectively. Design teams may also benefit from value advocates, but reports from the field show that in business-oriented settings value advocates met obstacles (Mander-Huits and Zimmer 2009). Their role has to be considered carefully, e.g., with respect to how much leadership they take and how other design team members receive such leadership. Consequently, another question for debate is whether the ethicist (or value advocate) should be a member of the design team or rather a member of a standards organization, which creates policies or regulations to be used by design teams?

## e-Social Sciences and Digital Humanities

Besides the inherent methodological issues above, new developments require attention from scholars. The ever-growing possibilities of receiving, storing, and managing immense amounts of data from diverse sources (e.g., mobile devices, RFID, cameras, wireless sensor networks or software logs) have led to recent trends

commonly referred to as "big data" and "data mining." Within social sciences economics, or humanitiies new fields of study are emerging called e-Social Sciences or digital humanities where the use of "mathematical models in search of recurrent patterns are seen ... as being of equal value compared to the traditional cycle of hypothesis formulation, observation, testing, validation or falsification and hypothesis reformulation" (van den Hoven et al. 2012). Companies, governments, and (research) institutions are equally interested in the benefits of big data, such as analyzing market trends, preventing diseases, or fighting crime. Institutional as well as policy design will be based on the knowledge retrieved from data mining with computer simulations often serving as test-beds (van den Hoven et al. 2012). Despite the potential benefits, big data also poses severe risks for privacy. Moreover, scholars have begun to provoke critical reflection and discussion around the assumptions, biases, and practices of big data science. Most notably, Boyd [sic] and Crawford (2011) have pointed to methodological issues (e.g., the definition of knowledge or objectivity in the data) as well as issues of data protection, epistemic responsibilities, and justice.

## Emerging ICTs

As we have seen in this chapter, ICTs have important (positive or negative) effects on human lives. Therefore, it is important to understand the ethical issues associated with their widespread use early on in the development process – not only to ensure value-sensitive functionality but also to inform policy makers and to increase the adoption of technologies with positive potential. For ethics or moral philosophy, it is, however, difficult to analyze technologies that are not yet implemented. Furthermore, "whatever provisionary understanding we manage to achieve is often outpaced by rapid change and new developments, making social and political interventions difficult" (van den Hoven et al. 2012).

Especially due to the complexities and pervasiveness of modern ICT, the diverse stakeholders and (sometimes even unpredictable) contexts of use, new ICTs, artifacts, applications, and functionalities will emerge that are unlikely to be predictable. Ethical analysis of emerging ICTs has to deal with several issues: (1) interpretive flexibility (i.e., technology being constituted by use leading to diverse use cases), (2) epistemological questions about how to get to a plausible account of the future (to analyze), and (3) an infinite number of possible ethical problems due to many possible futures (Stahl et al. 2010).

## Conclusions

In this chapter we have discussed diverse ethical implications of ICT and approaches to design for values in ICT. Due to the rapid development and widespread use of ICT in many diverse contexts, ranging from high-risk domains to the home environments, the stakeholders and values at stake are many.

Over time scholars from different disciplines including philosophy, social sciences, design, computer science, and human-computer interaction have shifted their focus from technical aspects of ICT to usability issues and recently to values implicated by ICT design. While some provide studies of sociocultural aspects of ICT use, others (most notably value-sensitive design) have provided concrete methods to account for values in ICT. We have given an overview of these approaches and methods as well as three concrete examples of values in ICT design.

As a fast-growing field, ICT will continue to pose new ethical questions that require new methods and innovate designs in the future. Some of the trends have been outlined above, but with the endless capabilities of computers, it is hard to foresee where developments will take us. Important, however, is that developers of ICT are aware of their responsibilities when designing ICT that will impact human lives and that they join forces with other scholars to ensure value-sensitive design of new ICTs.

## Cross-References


- ▶ Design for the Value of Human Well-Being
- ▶ Design for the Value of Privacy
- ▶ Design for the Value of Responsibility
- ▶ Design for the Value of Trust
- ▶ Design for the Values of Accountability and Transparency
- ▶ Participatory Design and Design for Values
- ▶ Value Sensitive Design: Applications, Adaptations, and Critiques


## References


Alsheikh T, Rode JA, Lindley SE (2011) (Whose) value-sensitive design: a study of long-distance relationships in an Arabic cultural context. In: Proceedings of the ACM 2011 conference on computer supported cooperative work. ACM, Hangzhou, pp 75-84

Bødker S, Grønbæk K (1991) Cooperative prototyping: users and designers in mutual activity. Int J Man-Mach Stud 34(3):453–478

Borning A, Muller M (2012) Next steps for value sensitive design. In: Proceedings 34th ACM SIGCHI conference on human factors in computing systems, Austin, Texas, pp 1125–1134

Borning A, Friedman B, Davis J, Lin P (2005) Informing public deliberation: value sensitive design of indicators for a large-scale urban simulation. In: Proceedings of ECSCW, Paris, France, pp 449–468

Bowker G, Star SL (1999) Sorting things out: classification and Its consequences. MIT Press, Cambridge, MA

Boyd D, Crawford K (2011) Six provocations for big data. A decade in internet time: symposium on the dynamics of the internet and society. Available at SSRN: http://dx.doi.org/10.2139/ssrn.1926431

Bozdag E, van de Poel I (2013) Designing for diversity in online news recommenders. In: Technology management in the IT-Driven Services (PICMET), proceedings of PICMET'13, Portland, Oregon, pp. 1101–1106

Branham SM, Thieme A, Nathan LP, Harrison S, Tatar D, Olivier P (2014) Co-creating & identity-making in CSCW: revisiting ethics in design research. In: Proceedings of 36th ACM conference on computer supported cooperative work & social computing, Vancouver, Canada, pp 305–308

Cavoukian A (2009) Privacy by design: The 7 foundational principles. Information and Privacy Commissioner of Ontario, Toronto

Cockton G (2006) Designing worth is worth designing. In: Proceedings of 4th Nordic conference on human-computer interaction: changing roles, Oslo, Norway, pp 165–174

Crabtree A (2003) Designing collaborative systems – a practical guide to ethnography, Springer series: computer supported cooperative work. Springer, London

Denning T, Kohno T, Levy HM (2013) Computer security and the modern home. Commun ACM 56(1):94–103

Detweiler CA, Hindriks KV (2012) Value-sensitive design patterns for pervasive health care. In: Proceedings of IEEE international conference on pervasive computing and communications, Lugano, Switzerland, pp 908–913

Detweiler CA, Pommeranz A, Stark L (2012a) Workshop on methods to account for values in human-centered computing held in conjunction with the ACM SIGCHI conference on human factors in computing systems – CHI'12, Austin, Texas

Detweiler CA, Dechesne F, Hindriks KV, Jonker CM (2012b) Ambient intelligence implies responsibility. In: Bosse T (ed), IOS Press Amsterdam, Ambient intelligence and smart environments, Ebook vol 12, Agents and ambient intelligence, IOS Press Amsterdam, The Netherlands, pp 33–61

Detweiler CA, Harbers M, Hindriks K (2014) Value stories: putting values into requirements engineering. In: Proceedings of workshop on creativity in requirements engineering (CreaRE), Essen, Germany

Fitrianie S, Huldtgren A, Alers H, Guldemond NAA (2013) SmartTV platform for wellbeing, care and social support for elderly at home. In: Biswas J, Kobayashi H, Wong L, Abdulrazak B, Mokhtar M (eds). Inclusive society: health and wellbeing in the community, and care at home. Lecture notes in computer science, vol 7910, Springer Berlin Heidelberg, pp 94–101

Flanagan M, Howe DC, Nissenbaum H (2005) Values at play: design tradeoffs in socially-oriented game design. In: Proceedings of ACM SIGCHI conference on human factors in computing systems, Portland, Oregon, pp 751–760

Flanagan M, Nissenbaum H, Belman J, Diamond J (2007) A method for discovering values in digital games. In: Situated play, proceedings of DiGRA conference. http://www.digra.org/wp-content/uploads/digital-library/07311.46300.pdf. Retrieved on 25 Aug 2014

Friedman B (1995) "It's the computer's fault": reasoning about computers as moral agents. In: Proceedings of ACM SIGCHI conference on human factors in computing systems, Denver, Colorado, pp 226–227

Friedman B, Hendry DG (2012) The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In: Proceedings of ACM SIGCHI conference on human factors in computing systems, Austin, Texas, pp 1145–1148

Friedman B, Kahn PH Jr (2003) Human values, ethics, and design. In: Jacko J, Sears A (eds) The human-computer interaction handbook. Lawrence Erlbaum Associates, Mahwah

Friedman B, Nissenbaum H (1996) Bias in computer systems. Trans Inf Syst (TOIS), 14(3):330–347

Friedman B, Nissenbaum H (1997) Software agents and user autonomy. In: Proceedings of the first international conference on autonomous agents, Marina del Rey, CA, USA, pp 466–469

Friedman B, Khan PH Jr, Howe DC (2000) Trust online. Commun ACM 43(12):34–40

Friedman B, Kahn P, Hagman J, Severson RL, Gill B (2006) The watcher and the watched: social judgments about privacy in a public place. Hum Comput Interact 21:235–272

Friedman B, Freier NG, Kahn P, Lin P, Sodeman R (2008a) Office window of the future? Field-based analyses of a new use of a large display. Int J Hum Comput Stud 66(6):452–465

Friedman B, Höök K, Gill BT, Eidmar L, Sallmander Prien C, Severson RL (2008b) Personlig integritet: a comparative study of perceptions of privacy in public places in Sweden and the United States. In: Proceedings of 6th Nordic conference on human-computer interaction, Lund, Sweden, pp 142–151

Friedman B, Nathan LP, Kane S, Lin J (2011) Envisioning cards. University of Washington, Seattle. Available at: envisioningcards.com

Friedman B, Denning T, Kohno T (2013a) Security cards: a security threat brainstorming toolkit. University of Washington, Seattle. Available at: securitycards.cs.washington.edu. Retrieved on 25 Aug 2014

Friedman B, Kahn PH, Borning A, Huldtgren A (2013b) Value sensitive design and information systems. In: Doorn N, Schuurbiers D, van de Poel I, Gorman ME (eds) Early engagement and new technologies: opening up the laboratory, vol 16, Springer series: philosophy of engineering and technology. Springer, Dordrecht, pp 55–59

Gowans C (2012) Moral relativism. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (Spring 2012 edn). The metaphysics research lab, Stanford, CA. URL: http://plato.stanford.edu/archives/spr2012/entries/moral-relativism/ Retrieved on 25 Aug 2014

Holtzblatt K, Jones S (1993) Contextual inquiry: a participatory technique for system design. In Schuler D Namioka A (eds.), Lawrence Erlbaum Associates, Hillsdale, pp 177–210

Huldtgren A, Wiggers P, Jonker CM (2013) Designing for self-reflection on values for improved life decisions. Interact Comput 2013. doi:10.1093/iwc/iwt025

Huldtgren A, Ascencio G, Pohlmeyer A, Romero Herrera N (2014) AAL-technology acceptance through experience. In: Proceedings of pervasive health 2014, Oldenburg, Germany

Johnson DG (1985) Computer ethics, 1st edn. Prentice Hall, Englewood Cliffs

Kahn PH, Friedman B, Gill BT, Hagman J, Severson RL, Freier NG (2008) A plasma display window? The shifting baseline problem in a technologically-mediated natural word. J Environ Psychol 28(2):192–199

Kensing F, Madsen KH (1992) Generating visions: future workshops and metaphorical design. Lawrence Erlbaum Associates, Roskilde, pp 155–168

Kling R (1999) What is social informatics and why does it matter? D-Lib Mag 5(1):205–220

Koch SH, Proynova R, Paech B, Wetter T (2013) How to approximate users' values while preserving privacy: experiences with using attitudes towards work tasks as proxies for personal value elicitation. Ethics Inf Technol 15(1):45–61

Koepfler JA, Fleischmann KR (2011) Classifying values in informal communication: adapting the meta-inventory of human values for tweets. Proc Am Soc Inf Sci Technol 48(1):1–4

Kraemer F, van Overveld K, Peterson M (2011) Is there an ethics of algorithms? Ethics Inf Technol 13(3):251–260

Le Dantec CA, Poole ES, Wyche SP (2009) Values as lived experience: evolving value sensitive design in support of value discovery. In: Proceedings of 27th international ACM, SIGCHI conference on human factors in computing systems, Boston, USA, pp 1141–1150

Mander-Huits N, Zimmer M (2009) Values and pragmatic action: the challenges of introducing ethical intelligence in technical design communities. Int Rev Inf Ethics 10:1–7

Marshall CC, Shipman FM (2011) Social media ownership: using twitter as a window onto current attitudes and beliefs. In: Proceedings of the SIGCHI conference on human factors in computing systems, Vancouver, Canada, pp 1081–1090

Miceli M, Castelfranchi C (1989) A cognitive approach to values. J Theory Soc Behav 19:169–193

Miller J, Friedman B, Jancke G, Gill B (2007) Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system. Proc GROUP 2007:281–290

Morandini M, Nguyen D, Perini A, Siena A, Susi A (2008) Tool-supported development with tropos: the conference management system case study. In: Luck M, Padgham L (eds) Agent-oriented software engineering VIII, lecture notes in computer science, vol 4951. Springer, Berlin/Heidelberg, pp 182–196

Movius LB, Krup N (2009) US and EU privacy policy: comparison of regulatory approaches. Int J Commun 3:19
Napoli P (1999) Deconstructing the diversity principle. J Commun 49(4):7–34
Nathan LP, Klasnja PV, Friedman B (2007) Value scenarios: a technique for envisioning systemic effects of new technologies. In: CHI '07 extended abstracts on human factors in computing systems, ACM, pp 2585–2590
Nissenbaum H (1998) Values in the design of computer systems. Comput Soc 28(1):38–39
Nissenbaum H (2005) Values in technical design. In: Encyclopedia of science, technology and ethics. Macmillan, New York, pp lxvi–lxx
Norman DA, Draper SW (1986) User centered system design: new perspectives on human-computer interaction. Boca Raton, Florida, CRC Press
O'Neill E (2000) User-development cooperation in software development: building common ground and usable systems (ed: van Rijsbergen CJ). Springer, London
Oxford Dictionary (2014) http://www.oxforddictionaries.com/definition/english/information-technology. Retrieved 25 Aug 2014
Pariser E (2011) The filter bubble: what the internet is hiding from you. Penguin Press, New York
Pommeranz A, Detweiler C, Wiggers P, Jonker CM (2011) Elicitation of situated values: need for tools to help stakeholders and designers to reflect and communicate. Ethics Inf Technol 14 (4):285–303
Potts C (1999) Scenic: a strategy for inquiry-driven requirements determination. In: Proceedings of IEEE fourth international symposium on requirements engineering (RE'99), Limerick, Ireland, pp 58–65
Rokeach M (1973) The nature of human values. Free Press, New York
Rosson MB, Carroll JM (2002) Scenario based design. In: Jacko J, Sears A (eds) The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications. Erlbaum, Boca Raton, pp 1032–1050
Rosson MB, Carroll JM (2009) Scenario based design. Human-computer interaction. Boca Raton, pp 145–162
Sallai G (2012) Defining infocommunications and related terms. Acta Polytech Hung 9(6)
Sanders EB-N, Westerlund B (2011) Experience, exploring and experimenting in and with co-design spaces. In: Proceedings of NORDES'11, Helsinki, pp 1–5
Sanders C, Rogers A, Bowen R et al (2012) Exploring barriers to participation and adoption of telehealth and telecare within the Whole System Demonstrator trial: a qualitative study. BMC Health Serv Res Healthc Needs Demand 12:220
Schneider FB (ed) (1999) Trust in cyberspace. National Academies Press, Washington, DC
Schwartz SH, Bilsky W (1990) Toward a theory of the universal content and structure of values: extensions and cross-cultural replications. J Pers Soc Psychol 58:878–891
Sharkey A, Sharkey N (2012) Granny and the robots: ethical issues in robot care for the elderly. Ethics Inf Technol 14(1):27–40
Shilton K (2012) Values levers: building ethics into design. Sci Technol Hum Values 38 (3):374–397
Simonsen J (2009) The role of ethnography in the design and implementation of IT systems. Des Princ Pract Int J 3(3):253–264
Stahl BC, Heersmink R, Goujon P, Flick C, van den Hoven J, Wakunuma K, Rader M (2010) Issues, concepts and methods relating to the identification of the ethics of emerging ICTs. Available at: http://www.academia.edu/432392/Issues_Concepts_and_Methods_Relating_to_the_Identification_of_the_Ethics_of_Emerging_ICTs. Retrieved on 25 Aug 2014
EIDD Stockholm Declaration (2004) Retrieved on 25 Aug 2014 from http://www.designforalleurope.org/Design-for-All/EIDD-Documents/Stockholm-Declaration/
Thew S, Sutcliffe A (2008) Investigating the role of 'soft issues' in the RE process. In: Proceedings of 16th IEEE international requirements engineering conference, Barcelona, Spain, pp 63–66
van den Berg N, Schumann M, Kraft K, Hoffmann W (2012) Telemedicine and telecare for older patients – a systematic review. Maturitas 73(2):94–114

van den Hoven J, Helbing D, Pedreschi D, Domingo-Ferrer J, Gianotti F, Christen M (2012) FuturICT-The road towards ethical ICT. The European Physical Journal Special Topics. arXiv preprint arXiv:1210.8181, 214(1):153–181

van Wynsberghe A (2013) Designing robots for care: care centered value-sensitive design. Sci Eng Ethics 19(2):407–433

Woelfer JP, Iverson A, Hendry DG, Friedman B, Gill BT (2011) Improving the safety of homeless young people with mobile phones: values, form and function. In: Proceedings of ACM SIGCHI conference on human factors in computing systems, Vancouver, Canada, pp 1707–1716

Yoo D, Huldtgren A, Woelfer JP, Hendry DG, Friedman B (2013) A value sensitive action-reflection model: evolving a co-design space with stakeholder and designer prompts. In: Proceedings of the ACM SIGCHI conference on human factors in computing systems, Paris, France, pp 419–428

# Design for Values in Institutions

Seumas Miller

## Contents



**Abstract**

In this chapter, I examine the relationship between institutions and moral or ethical values (I use the terms interchangeably) and, in particular, the manner and extent to which such values are or, at least, ought to be part and parcel of the design of institutions. By institutions I mean organizations and systems of organizations. So a single business corporation is an institution in this sense, and so is a market-based industry comprised of corporations. When designing-in-values to institutions, three dimensions of institutions can be considered, namely, function, structure, and culture. Moreover, there are different (possibly crosscutting) levels: the macro-level (e.g., the industry as a whole), mezzo-level (e.g., a single organization), and the microlevel (e.g., an anti-corruption system within an organization). Further, there are at least six main sources of motivation to be accommodated, and potentially utilized, in the design process. These are formal sanctions (within a framework of enforced rules), economic incentives (especially within a competitive market), desire for status and reputation, desire

S. Miller (✉)
Australian National University, Canberra, Australia
e-mail: s.r.m.miller@tudelft.nl; seumas.miller@anu.edu.au



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_27

for control over one's own destiny and (in some cases) power over others, moral motivations, and a miscellaneous assemblage of psychosocial factors, e.g., status quo bias, overconfidence, desire to conform, and irrational desires. To illustrate and facilitate understanding of designing-in-value to institutions, I will discuss different features of a variety of quite diverse contemporary institutions. The institutions and design features in question are (respectively) (1) the design and construction of an entire organizational system from the ground up, namely, a compulsory retirement income system (a hybrid public/private sector institution); (2) the redesign and renovation of an anti-corruption system for existing police organizations (public sector institution); (3) the design and construction of a reputational index for organizations competing in a market as one element of a broad-based cultural-change process for the industries in question; and (4) the redesign of disclosure requirements for credit card pricing mechanisms.

**Keywords**
Institutions • Design for values • Anti-corruption system • Reputational indices

## Introduction

Sometimes the term "institution" is used to refer to simple social phenomena such as conventions, e.g., handshakes, and sometimes to complex social forms that are not organizations such as human languages or kinship systems. However, the concern in this chapter is only with institutions that are also organizations and/or systems of organizations.

Such organizations (or systems thereof) are complex social forms that reproduce themselves and include governments, police organizations, universities, hospitals, business corporations, markets, and legal systems. Moreover, social institutions in this sense are among the most important of collective human phenomena; they enable us to feed ourselves (markets and agribusinesses), to protect ourselves (police and military services), to educate ourselves (schools and universities), and to govern ourselves (governments and legal systems). In short, institutions have purposes or functions.

Institutions also have structure and vary greatly in this regard. Compare, for example, the hierarchical top-down structure of military organizations with the flat democratic structures typical of amateur sporting clubs.

The third main dimension of institutions is culture, the "spirit" or informal set of attitudes that pervades an organization and which might reinforce or negate the more formal requirements of the organization. An example of the latter is the culture in certain police organizations which protects those engaged in corruption rather than exposing them.

A further feature of institutions is their use of technology; indeed, technology is in part constitutive of most, if not all, modern institutions. Some technologies have specialized functions and are only used by certain kinds of institutions, e.g., weaponry used by military institutions. Other technologies have generic functions

and are used by virtually all institutions. Consider, for example, the use of the Internet by modern organizations. This close relationship between institutions and technology has led to the use of terminology such as the term "socio-technical systems." Terminology aside, the design of institutions typically involves close attention to relevant technology; after all, institutions shape and are shaped by technology. However, my concern in this chapter is with institutions per se, and my periodic references to, and descriptions of, technology need to be seen in this context.

Viewed synchronically, institutions are multi-level collective entities. That is, they have different, and possibly crosscutting, levels. For example, the global banking sector can be viewed at the macro-level: the industry as a whole. At this level there are multiple banks in market-based competition with one another. However, at what might be referred to as the mezzo-level, there is simply a single organization: one bank. Further, at the microlevel there are various sub-organizational units, such as a single bank's fraud and anti-corruption unit. These levels are crosscutting from an institutional design perspective since, for example, a regulatory change, such as the requirement to segregate the investment arm from the retail banking arm, might affect all the banks but do so only in respect of the internal structure of each.

Viewed diachronically, institutions can be thought of as constantly evolving entities, e.g., as increasing or decreasing in size, capacity, and so on. An important distinction to be made here is between institutions which have evolved more or less unconsciously and those which have been consciously designed. Legal systems are an example of the former, whereas the modern limited liability corporation is an instance of the latter. Naturally, the boundary between these two kinds of institution is vague. Moreover, most modern institutions have at least some organizational components which have been consciously designed, e.g., information and communication technology (ICT) service units. Further, some institutions can be designed and constructed wholly anew from the ground up; others cannot be. An example of the latter might be the criminal justice system. Here the image of Theseus's ship comes to mind. While a ship is at sea, parts of it can be changed piecemeal, but the ship can be demolished and rebuilt from the ground up. On the other hand, a ship redesigned and rebuilt in a piecemeal fashion may end up being radically different at the end of the process from what it was at the start. Consider, for example, the institutional redesign and rebuilding of the governance structure in post-apartheid South Africa.

Among other things, a normative theory of institutions specifies what the purpose or function of particular types of institution *ought to be*, as opposed to what in fact it is. Enron, for example, apparently had the de facto institutional purpose of enriching its CEO and other senior officers, but this was surely not what its institutional purpose ought to have been.

One normative theory of social institutions is based on an individualist theory of joint action (Miller 2010). Put simply, on this account the organizations or systems of organizations are ones that provide collective goods by means of joint activity. The collective goods in question include the fulfillment of aggregated moral rights,

such as needs-based rights for security (police organizations), material well-being (businesses operating in markets), education (universities), governance (governments), and so on.

Self-evidently, institutions in this sense have a multifaceted ethico-normative dimension, including a moral dimension. Moral categories that are deeply implicated in social institutions include human rights and duties, contract-based rights and obligations, and, importantly on this teleological normative account, rights and duties derived from the production and "consumption" of collective goods. (This is not the economists' notion of non-rival and non-excludable goods.)

Collective goods of the kind in question have three properties: (1) they are produced, maintained, or renewed by means of the *joint activity* of members of organizations or systems of organizations, i.e., by institutional actors; (2) they are *available to the whole community* (at least in principle); and (3) they *ought* to be produced (or maintained or renewed) and made available to the whole community since they are desirable goods and ones to which the members of the community have a (institutional) *joint moral right*.

Such goods are ones that are desirable in the sense that they ought to be desired (objectively speaking), as opposed to simply being desired; moreover, they are either intrinsic goods (good in themselves) or the means to intrinsic goods. They include, but are not restricted to, goods in respect of which there is an institutionally prior moral right, e.g., security.

Notice that on this account of institutions institutional design will consist in large part in designing and establishing institutions that realize collective goods and, in the case of existing institutions, redesigning them to ensure that they continue to provide such goods in circumstances in which they are failing to do so. Notice also that this normative account is consistent with a description of the ontological genesis of institutions, according to which it is a mix of unconscious evolutionary social processes and conscious design.

Whether one accepts this normative theory or some other, at the most fundamental level, designing-in-values to an institution must be done in the context of some account of the institutional function or purpose of the institution in question. Moreover, to some extent this purpose or function will determine what is an appropriate structure and culture. After all, structure and culture ought to facilitate institutional purpose. Perhaps a hierarchical structure is necessary if military organizations are to realize their institutional purposes of successfully waging war. On the other hand, top-down hierarchical structures may not be conducive to academic work and, therefore, ought not to be imposed on universities.

We need to distinguish between the institutional purpose of a single organization and that of the system of organizations of which it is a part. An obvious example here is that of market-based organizations. Particular market actors come and go; indeed, according to the normative theory of the competitive market, this is desirable. However, the institution, the market-based industry, remains. According to one normative account of the economic mechanism of the "invisible hand," the institutional purpose of markets is to maximize utility. On a contrary account, it is to produce an adequate quantum of reasonably priced goods or services of

reasonable quality. Importantly, on both accounts it is the set of business firms taken as a whole which realizes the institutional purpose rather than any one organization considered on its own.

We need to make further distinctions between market-based and nonmarket-based institutions, e.g., governments and most police organizations, and also between both of these and hybrid institutions – part market based, part nonmarket based. An instance of the latter discussed below is that of a compulsory retirement savings scheme.

Market-based institutions are designed with a view to the incentive of material gain: money, in short. Market will work effectively only if the structure of economic incentives is appropriately designed. By contrast, criminal justice institutions are designed in the light of the motivating force of enforceable laws, rules, and regulations. Institutional actors will comply because if they don't sanctions will follow. This is, of course, a simplification. The design of most, if not all, contemporary institutions utilizes financial incentives as well as rules backed by formal sanctions. Moreover, there are other pervasive drivers in play among most institutional actors. One of these is the desire for control of over one's own destiny (autonomy) and (perhaps) control over others (power). Another is reputation or status. The latter is probably more important to academics than, for example, financial reward. And there are moral motivations. Institutions depend heavily on institutional actors having a moral sense and acting on it. If one is inclined to doubt this, then consider the destructive forces that are unleased when the members of a given institution abandon basic moral precepts and ignore their institutional and moral duties to one another and to those they are employed to serve. Corrupt government and other officials in impoverished African states, such as the Central African Republic, are a case in point. Finally, there is a miscellaneous assemblage of psychosocial factors, e.g., status quo bias and overconfidence, that affect institutional actors and, therefore, can facilitate or impede institutional design. Many of these have been identified and described by Richard Thaler and Cass Sunstein in their influential book, *Nudge* (2008).

## Designing and Building an Institution from the Ground Up: Compulsory Retirement Income Systems

Superannuation is the name of Australia's compulsory retirement income system. Superannuation differs from some other institutions in that it has been intentionally designed by Australian policymakers to serve a particular purpose (provide for the financial needs of retirees) and do so by a specific means (compulsory savings for the workforce). In short, the collective good which Australia's compulsory retirement income system exists to provide is an aggregate of needs, namely, the aggregate of financial needs of Australian retirees. Recently, the Cooper Review (Cooper 2012) identified a raft of deficiencies in the current Australian superannuation system pertaining to its efficiency and effectiveness and has sought to redesign it in a variety of ways. However, in doing so it has left its basic structure intact.

While superannuation has an important moral purpose, a defining characteristic of the system is that the moving parts are outsourced almost entirety to the private sector. The majority of Australians are members of institutional superannuation funds that are required to be set up as trusts to offer the members the protection trust law has traditionally provided to beneficiaries. The trustees of most funds outsource the investment management and administration to third parties, who may or may not be connected to the fund. Typically, upon taking a new job, employees are automatically enrolled as a member of their new employer's default fund unless they elect to have their superannuation contributions directed to another fund.

Accordingly, superannuation is a hybrid public/private sector institution and has been specifically designed as such. Thus, the division between market-based and nonmarket-based institutions is not necessarily a strict dichotomy; there is also the possibility of hybrid institutions: institutions that utilize the mechanism of the market but are not wholly market based. The Australian compulsory retirement savings system is a case in point, but there are many others. Needless to say, by the lights of our normative teleological account of social institutions, whether an institution ought to be wholly market based, nonmarket based, or a hybrid of both is a matter to be settled by recourse to collective goods: which of these three models most efficiently and effectively enables the collective good definitive to the social institution in question to be produced.

From the perspective of designing-in-values, there are some interesting aspects of the superannuation scheme, aside from its basic institutional purpose of providing for the financial needs of retirees. The first point is that the scheme is compulsory; so the freedom of Australian workers is being infringed in the service of a larger purpose. An important moral question here is whether or not individual freedom should be overridden by their future well-being. So there is an element of paternalism here. The second point is that within the constraint of compulsory saving, the scheme does attempt to maximize freedom by allowing for the possibility for self-managed funds. Australian can choose whether they are in charge of their own superannuation fund or they consent to their funds being managed by someone else.

Regarding the second point, surely maximizing freedom is morally right, other things being equal. What of the first point? Here it is important to note that if workers were not forced to save to provide for *their own* retirement, others would be forced to provide for them instead. This would surely be unfair. Therefore, the argument against compulsory savings based on the infringement of the freedom of workers collapses.

## Redesigning an Anti-corruption System for Police Organizations

According to one teleological normative theory of police organizations, the protection of aggregate moral rights (a collective good) is the central and most important moral purpose of police work, albeit a purpose whose pursuit ought to be constrained by the law. So while police institutions have other important purposes

that might not directly involve the protection of moral rights, such as to enforce traffic laws or to enforce the adjudications of courts in relation to disputes between citizens, or indeed themselves to settle disputes between citizens on the streets, or to ensure good order more generally, these turn out to be purposes derived from the more fundamental purpose of protecting moral rights, or they turn out to be (non-derivative) secondary purposes. Thus, laws against speeding derive in part from the moral right to life, and the restoring of order at a football match ultimately in large part derives from moral rights to the protection of persons and of property. On the other hand, service of summonses to assist the courts is presumably a secondary purpose of policing.

While police ought to have as a fundamental purpose the protection of moral rights, their efforts in this regard ought to be constrained by the law. In so far as the law is a constraint – at least in democratic states – then this view accommodates "consent" as a criterion of legitimacy for the police role. However, on this view legality, and therefore consent, is only one consideration. For police work ought to be guided by moral considerations – namely, moral rights – and not simply by legal considerations.

An important consequence of this account is that police organizations need to be designed in such a way that they have a degree of independence of government. For one thing, governments might engage in illegality or the violation of the rights of their citizenry. For another, the primary commitment of police organization must be to the law (and not to "men," so to speak). On the other hand, police organizations need to be responsive to the democratically elected government (and other democratic bodies). This gives rise to difficult issues of institutional design of values, specifically, independence versus responsiveness.

Historically, high levels of police corruption have been a persistent tendency in police services throughout the world. Accordingly, designing anti-corruption systems for police organizations has been a major focus of attention for governments and police organizations alike. Some of the designed-in-features are as follows:

- A stringent vetting process to prevent corrupt and other inappropriate applicants from entering the organization
- Intelligence gathering, risk management, and early warning systems for at-risk officers, for example, officers with high levels of complaints
- Proactive anti-corruption intervention systems, for example, targeted integrity testing
- External oversight by an independent, well-resourced body with investigative powers

The first point to be made is that the institutional purpose of an anti-corruption system is a moral purpose; so designing an anti-corruption system is designing-in-ethics at a fundamental level. However, the above features give rise to a variety of ethical issues. A vetting process which excludes applicants because of the level of ethical risk they pose might be regarded as unfair. Perhaps it is in some cases, e.g., persons who have criminal associates but who are not themselves known to have

engaged in any criminal activity. However, arguably, society's interest in relatively corruption-free police service overrides fairness to some individuals in respect of job opportunities.

External independent oversight is surely unexceptionable and has been found to be effective against police corruption. Interestingly, however, internal investigation units have been continued to work to some extent in tandem with these external bodies. This arrangement enables the police organization itself to continue to take a high degree of responsibility for its own unethical officers. So the design accommodates both considerations: "owning your own corruption" but also being subject to external oversight and, if necessary, intervention.

Recent developments in ICT have enabled the development of, for example, early warning systems based on a complex mix of risk factors. This is a good example of technology deployed in the service of institutional purpose.

Targeted integrity testing has been found to be an effective means of identifying and removing corrupt officers. However, it has also generated intense ethical debate, given that integrity testing involves trapping or "stinging" an unsuspecting officer into performing a corrupt action which, but for the trap, he would not have performed. That is, he would not have performed that particular token act. Of course, if the trap is ethically well designed, then the trapped officer was, nevertheless, in the habit of performing similar corrupt actions.

## Institutional Design and Mobilizing Reputation: Reputational Indices

Business organizations and markets have evolved over hundreds of years. However, they have done so in the context of frequent government interventions of redesign, notably by means of regulation, and in the service of collective goods.

In the case of business organizations and markets, the collective goods include (i) the coordination of buyers and sellers of goods and services and (ii) a quantum of a product or service sufficient to meet the relevant aggregate needs of the population in question. Here Adam Smith's invisible hand mechanism is salient. The outcome (a collective good) of the workings of the invisible hand is the ultimate purpose (collective end) of this institutional mechanism (e.g., an adequate supply of houses, of auditing services, of retirement savings, etc.); profit maximization is the proximate end. Moreover, arguably the quantum of goods or services in question ought to be reasonably priced and of reasonable quality.

As we have seen, to realize their institutional purposes, market-based industries have relied predominantly on economic incentives and enforceable rules. However, I suggest there is another motivational source that might be utilized, namely, reputation. Naturally, some groups and organizations are more sensitive to reputational loss and the possibility of reputational gain than others. Corporations and professional groups in the financial services sector, including bankers and auditors, are very sensitive to reputational loss. Those entrusted to make prudent decisions with other people's money are inevitably heavily dependent on a good reputation,

similarly, for those entrusted to provide independent adjudications in relation to financial health.

When a high professional reputation is much sought after by members of an occupational group or organization and a low one to be avoided at all costs, there is an opportunity to mobilize this reputational desire in the service of promoting ethical standards. Here the aim is to ensure that reputation aligns with actual ethical practice, i.e., that an organization's or group's or individual's high or low reputation is deserved. One means to achieve this is the reputational index. Such an index could be constructed whereby an ethics audit awards scores in relation to specific ethical standards. Let me now sketch the broad outlines of such a reputational index.

Deserved reputation can provide an important nexus between the self-interest of corporations and professional groups, on the one hand, and appropriate ethical behavior toward consumers, clients, and the public more generally, on the other hand. More specifically, the deserved reputation of, say, financial service providers, be they corporations or professional groups, can provide such a nexus. Here there are three elements in play: (i) reputation, (ii) self-interest, and (iii) ethical requirements, such as particular ethical standards such as compliance with technical accounting standards and avoidance of specific conflicts of interest, but also more general desiderata such as client/consumer protection. The idea is that these three elements need to interlock in what might be called a virtuous triangle.

First, reputation is linked to self-interest; this is obviously already the case – individuals, groups, and organizations desire high reputation and benefit materially and in other ways from it. Second, reputation needs to be linked to ethics in that reputation ought to be deserved; reputational indices are a possible means to help to achieve this. Third, and as a consequence of the two already mentioned links, self-interest is linked to ethics; given reputational indices that mobilize reputational concerns, it is in the self-interest of individuals, groups, and firms to comply with ethical standards (which are also professional standards). Here I reassert that self-interest is not the only or necessarily the ultimate motivation for human action; the desire to do the right thing is also a powerful motivator for many, if not most, people. Accordingly, the triangle is further strengthened by the motivation to do right.

In recent years, the notion of a Reputation Index has gained currency in a number of contexts, especially in business and academic circles. The term seems to have a number of different senses. Sometimes it is used to describe a way of measuring the reputation that an organization actually has, since reputation exists, so to speak, in the eye of the beholder. Actual reputation does not always match deserved reputation. Accordingly, sometimes the term is used to describe a way of calculating the performance of an organization on the basis of which its reputation should be founded.

The first step in the process is to determine a way of accurately measuring the ethical performance of individual or organizational members of occupational and industry groups; this is an ethics audit.

Here I stress the importance of *objective* measures of ethical performance. The latter might include such things as results of consumer satisfaction surveys; gross

numbers of warranted complaints and trends thereof; numbers of disciplinary matters and their outcomes; and outcomes of financial, health, and safety audits (e.g., regarding electronic crime and corruption vulnerabilities). It would also include the existence of institutional processes established to assure compliance with ethical standards, e.g., codes of ethics and conduct, financial and other audit processes, ethics committees, complaints and disciplinary systems, fraud and ethics units, ethical risk assessment processes, ethics and compliance officers, and professional development programs in ethics.

Here I note that while some of these institutional systems and processes might be legally required – and, indeed, some are under various pieces of legislation – this is by means the case for all them. In short, while reputational indices include some indicators of compliance with those ethical standards (and associated processes of assurance) that are enshrined in law, they also include indicators of adherence to ethical standards that are above and beyond what is legally required.

In addition to the ethics audit itself, there is a need for a process that engages with ethical reputation. Since ethical reputation should reflect the findings of the ethics audit, an ethical reputation audit should drive the relationship between de facto ethical performance (in effect, the deserved reputation) and actual reputation for ethical performance. The way to achieve this is by the participation of as many occupational members and industry organizations as possible in ethics audits and by the widespread promulgation of the results of their de facto ethical performance (as determined by the ethics audit), including in the media. Naturally, the results promulgated could be more or less detailed; they could, for example, simply consist in an overall rating as opposed to a complete description of the ethics audit results.

## Record, Evaluate, and Compare Alternative Prices (RECAP)

In addition to the basic motives we have utilized in the above cases of designing-in-values (e.g., desire for a good reputation in the case of reputational indices), there are a miscellaneous assemblage of psychosocial factors which influence our behavior and which can be likewise utilized, or at least taken into account, in designing-in-values to institutions. Some of these psychosocial factors are motives, others are cognitive dispositions, and still others are character traits. They include status quo bias, overconfidence, the desire to conform, and various irrational impulses and desires which tempt us and weaken our rational wills, e.g., desires for instant gratification.

Richard Thaler and Cass Sunstein have developed a theoretical design posture – which they refer to colloquially as "nudge" – based in part on a recognition and understanding of these psychosocial factors (Thaler and Sunstein 2008). According to Thaler and Sunstein, the proposition that humans typically act in accord with their rational self-interest is false. Firstly, humans inevitably make choices in institutional and other settings which structure their behavior in various ways, including ways contrary to their individual self-interest. Secondly, their behavior

is heavily influenced by these abovementioned psychosocial factors which tend to undermine rationally self-interested action. Consider credit card pricing schemes (Thaler and Sunstein 2008, p. 93). On the one hand, modern economic life makes it difficult to function without a credit card; credit cards are a part of the financial architecture. On the other hand, credit card pricing schemes are very complex and often tailored to the commercial interests of the credit card providers rather than to the interests of the credit card users. Moreover, consumers do not take time to calculate how much each of the available credit card options costs and make a decision on that basis. Indeed, most credit card holders do not even know how much their credit card costs them, let alone how it compares cost-wise with other credit cards. This might not be very important were it not for the fact that many credit card users get into serious financial trouble by running up large amounts of credit card debt which they then struggle to repay (Thaler and Sunstein 2008, pp. 142–143).

What is to be done? Thaler and Sunstein argue, firstly, that the financial or other kind of architecture can often be adjusted or redesigned in various ways so as to influence choices; in their credit card example, it is the choices of credit card users that are in question. Someone who thus redesigns (in this case, financial) architecture is a choice architect (in their terminology). An important category of choice architects is government authorities who introduce regulations to influence behavior. Secondly, they argue that if a choice architect (say, the government) redesigns, for example, credit card pricing mechanisms by means of regulations in order to benefit consumers in significant ways, then this may well be justified, notwithstanding that it is paternalistic.

This is not the place to embark on an extended moral analysis of paternalism versus libertarianism. Nevertheless, it is worth pointing out that Thaler and Sunstein are advocating a relatively mild form of paternalism, since choices are not eliminated or even rendered excessively costly; rather they are recommending "nudging." They state: "A nudge... is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic options" (Thaler and Sunstein 2008, p. 6). Let us now turn to an example of institutional redesign of the "nudge" variety, namely, RECAP.

RECAP stands for record, evaluate, and compare alternative prices. RECAP is essentially a simplified disclosure requirement for sellers of products, the usage and relative costs of which "challenge" consumers leading to less than fully rational market behavior. Let us see how this might work with credit cards. Thaler and Sunstein suggest (Thaler and Sunstein 2008, p. 143) that credit card providers be required to send an annual statement to each user that lists and totals all the fees that the user has incurred over the year. This would give users a clear idea of how much they are paying and what for. Moreover, the statement should be in electronic form to enable users to compare what they are paying with what other credit card providers charge. In short, the RECAP device would facilitate much better choice, indeed, more rational choices, on the part of credit card users and may well enable many to avoid falling into a credit card debt trap.

## Conclusion: Values and the Design of Institutional Reality

In this chapter, the relationship between institutions and moral or ethical values has been examined and, in particular, the manner and extent to which such values can and ought to be designed into institutions. We identified six main sources of motivation to be accommodated, and potentially utilized, in the design process. These are formal sanctions (within a framework of enforced rules), economic incentives (especially within a competitive market), desire for status and reputation, desire for control over one's own destiny and (in some cases) power over others, moral motivations, and a miscellaneous assemblage of psychosocial factors, e.g., status quo bias, overconfidence, desire to conform, and irrational desires. To illustrate and facilitate understanding of designing-in-value to institutions, we discussed (1) the design and construction of an entire organizational system from the ground up, namely, a compulsory retirement income system; (2) the redesign and renovation of an anti-corruption system for existing police organizations; (3) the design and construction of a reputational index for organizations competing in a market as one element of a broad-based cultural-change process for the industries in question; and (4) the redesign of disclosure requirements for credit card pricing mechanisms.

There are a number of general conclusions to be drawn from the above. First, and perhaps most obviously, there are a diverse range of psychosocial factors that can be mobilized in the service of designing-in-values. These include the so-called sticks and carrots, notably enforceable rules and economic benefits, respectively. But they also include the desire to do what is morally right or worthwhile. Moreover, there are a range of less obvious psychosocial factors, including irrational desires and cognitive dispositions, that can facilitate design or which need to be otherwise accommodated by institutional designers.

Second, in relation to these psychosocial factors, it is not necessarily a matter of relying solely on one motive, e.g., enforceable rules or desire for economic benefit, albeit in any given case one or more of these motives might predominate. Rather it is typically a matter of devising an institutional arrangement which is underpinned by an integrated structure of motives. This is explicitly so in the case of the triangle of virtue used in the design of reputational indices. But it is also true of the compulsory retirement savings scheme and the anti-corruption systems for police organizations. The latter relied on a desire to do what is morally right as well as on enforceable rules. The former involved enforceable rules (deterring noncompliance), economic incentives ("carrots"), and moral considerations, such as fairness. Again, the disclosure requirements for credit card pricing mechanisms relied on enforceable rules (legal requirement to simplify and disclose pricing mechanism) as well as removing a cognitive-based impediment to rational behavior (complex pricing mechanisms).

Third, when designing-in-values to institutions, three dimensions of institutions can be considered, namely, function, structure, and culture. It should be kept in mind that on the teleological account of institutions elaborated above, function or purpose (collective good) gives, or ought to give, direction to structure and culture. Both structure and culture ought to facilitate institutional purpose. There is a

tendency for institutional designers to focus on structure to realize purpose and ignore culture, perhaps because culture is more nebulous and less tangible than structure. However, as we saw in the case of the reputational indices and the anti-corruption system for police organizations, culture is often critical and, therefore, in need of an institutional designer's attention.

Fourth, given the logical and moral priority of institutional purposes over structure and culture, a fixation with particular institutional structures can be unhelpful. Consider those whose first instinct is to opt for market-based solutions or, alternatively, public sector agencies. Often these responses are ideologically based rather than evidence based. Indeed, as we saw in the case of the compulsory retirement savings scheme, the reputational indices, and the disclosure scheme for credit card pricing mechanisms, the best design options may involve designing hybrid models involving both market-based and public sector components (or, at least, significant regulatory intervention to facilitate institutional purpose).

Fifth, at the macro-level of institutions, that is, systems of organizations, it may be difficult or even impossible in practice to design from the ground up. Accordingly, the design process may inevitably have to restrict itself to piecemeal redesigning of, or "renovating," particular features of the institution. This is perhaps the case with institutions such as legal systems or important market-based industries that have evolved over long periods of time. Nevertheless, important institutions, such as the compulsory retirement savings system, can be designed from the ground up. Here there is a need to distinguish political feasibility from institutional possibility or desirability.

## Cross-References


- ▶ Design for the Value of Regulation
- ▶ Design for the Values of Democracy and Justice
- ▶ Design for Values in Economics


## References


Cooper J (2012) The "Cooper" review. http://www.supersystemreview.gov.au/content/content.aspx?doc=html/final_report.htm. Accessed 21 May 2012

Fung A (2003) Survey article: recipes for public spheres: eight institutional design choices and their consequences. J Polit Philos 11(3):338–367

Goodin RE (ed) (1998) The theory of institutional design. Cambridge University Press, Cambridge

Miller S (2010) The moral foundations of social institutions: a study in applied philosophy. Cambridge University Press, New York

Searle JR (1995) The construction of social reality. Penguin, London

Thaler R, Sunstein C (2008) Nudge: improving decisions about health, wealth and happiness. Yale University Press, New Haven

Van den Hoven J, Miller S, Pogge T. Designing-in-ethics: technology and institutions. Cambridge University Press, New York (forthcoming)

# Design for Values in Nanotechnology

Urjan Jacobs and Marc de Vries

## Contents



**Abstract**

Applications of nanotechnology have the potential to raise fundamentally new ethical questions. Nanotechnology is an enabling technology and therefore a whole array of moral values is at stake. We investigate these values by differentiating with respect to specific applications. We will argue that in the short term, nanotechnology does not pose novel value-laden socio-technical issues, but has the potential to enhance or provide opportunities to address existing issues. We will describe three different attempts to provide a design for safety or sustainability approach, which are specific for nanotechnology. In the long term,

U. Jacobs (✉) • M. de Vries
TU Delft, Delft, The Netherlands
e-mail: j.f.jacobs@tudelft.nl; m.j.devries@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_29

nanotechnology does raise new ethical questions, especially with the blurring of category boundaries. Since the current debate on long-term developments is mainly technology assessment oriented in nature, we will suggest how these outcomes can be used for a more design-oriented approach.

**Keywords**
Cybernetic organism • Enabling technology • Human enhancement • Nanoethics • Nanoscale titanium dioxide

## Introduction

Nanotechnology is an intriguing technology, not in the least because of the ethical questions it evokes. Nanotechnology is the manipulation of structures at the nanometer scale (one nanometer is a billionth of a meter). This is only a rough description of what nanotechnology entails and a broader discussion on the definition will be provided in section "Description of Nanotechnology". Much of nanotechnology is still in the laboratory phase, and for that reason the term nanoscience is often more appropriate than nanotechnology. Nonetheless, some results are already on the market (first-generation nanotechnology) and others are about to be realized commercially. These current applications of nanotechnology may not give rise to fundamentally new ethical questions, but the wide variety of applications and possibly far-reaching consequences have led to the situation that the design and development of nanoproducts is surrounded by social debates that are often organized and facilitated by governments. As the development of nanotechnology is influenced by a variety of aspects, nanoethics is complicated and involves knowledge from a variety of disciplines (Vries 2006, 2008). In this chapter we will analyze what kind of ethical issues are at stake with the current developments and discuss some first attempts to provide a "Design for Values" approach specific for nanotechnology.

There are also long-term developments with possibly very important impacts that are already discussed now, in spite of the fact that speculation is involved in such debates (Grunwald 2010; Nordmann 2007). In the long-term in particular, new ethical issues seem to emerge. The new domain of synthetic biology, for instance, raises new questions about boundaries between natural and artificial and ethical questions related to that (for instance, are natural and artificial "life" equally worthy to protect?). Therefore, short-term and long-term developments will be discussed separately. The long-term debates often have a technology assessment-oriented nature: possible effects are studied or imagined, and based on the outcomes of that, a general assessment is made of whether or not we should develop such an application. In this chapter we will use literature of that kind, but also seek a more design-oriented approach in which we will ask the question what role values could and should play in the development of those applications. Of course, the outcomes of the technology assessment type of studies can be used for such design-oriented considerations as they provide clues of what is in line with certain values and what is not.

One of the interesting aspects of nanotechnology is that several authors have claimed that it raises new ethical issues (Ferrari 2010; Preston et al. 2010; McGinn 2010). It can always be debated whether or not an ethical question is truly novel or not. As we will see, nanoethics is certainly not fundamentally different from ethics in other technological domains. But particularly in the long-term expectations, we do see complications for Design for Values.[1] As Poel (2008) argues, we should not only focus on seemingly new ethical issues as we may then overlook other important issues. He also makes the point that important ethical issues may only become clear during the further development of nanotechnology (Poel 2008). In establishing values we often refer to certain categories that we are used to. Intuitively we divide in living versus nonliving, healthy versus ill, natural versus artificial, and the like and value certain categories over others. For instance, we may opt for an ethical stance in which natural is better than artificial (e.g., in the case of food) or living things are more worthy of protecting than nonliving things. Certain applications in nanotechnology tend to confuse the boundaries between such categories (Swierstra et al. 2009; Verbeek 2009). That creates a problem when assessing values. Thus, Design for Values can become problematic, as it is not clear what values are at stake or how they relate to certain categories.

In this contribution we will give an overview of nanotechnology, before we will analyze the ethical issues that are at stake in the short- and long-term development of nanotechnology. We will then give an overview on three preliminary attempts to provide a "Design for Values" approach that are specific for short-term nanotechnology development; we will also discuss approaches for the longer term. To provide ample context to the approaches, the ethical issues with current application of nanoparticles in sunscreens and the long-term application of cyborgs are discussed. We will end the contribution by giving suggestions for further work as well as drawing conclusions.

## Nanotechnology

Within a decade, nanotechnology has become a major technological theme across most scientific and engineering disciplines. Especially since the start of the US-based National Nanotechnology Initiative (NNI) in 2000, nanotechnology captured the imagination of various stakeholders. Governments all over the world have launched and promoted nanotechnology programs, initiatives, and business alliances to benefit from the identified economic potential that nanotechnology promises to bring as well as to keep up with scientific and technological advances elsewhere. The almost unprecedented technological movement on a global scale has been stimulated by promises of a "next industrial revolution" (Committee on Technology 2000). Nanotechnology thus may appear like a creation of politicians

[1]Here, we take the term Design for Values in a sense that is wider than "value-sensitive design"; see Hoven and Manders-Huits (2009).

given these strong political efforts by governmental funding and stimulation. Nonetheless, products with nanosized materials as well as components are currently being designed, produced, and used. The application of nanotechnology will likely grow further as spending in nanotechnology-related R&D increases (Malanowski et al. 2006; Rensselaer 2004).

## Description of Nanotechnology

Nanotechnology works in the area between isolated molecules and larger solids, regularly referred to as the size range of 1–100 nm. Phenomena occur in this transient area, which are not observed on molecular nor on macroscopic objects. Nanotechnology can be used in numerous application areas, such as agriculture, chemical industry, construction, cosmetics, energy, health care, information technology, textiles, and transport (Malanowski et al. 2006). Currently, nanomaterials are utilized in various commercial products already on the market, including antimicrobial wound dressings, antifog layers, food packaging, chemical catalysts, multimedia data recorders, cosmetics, LED-based lighting, diode lasers, low-friction coatings, microelectronics, and sunscreens. The Project on Emerging Nanotechnologies of the Woodrow Wilson International Center for Scholars and the Pew Charitable Trusts keeps an inventory of manufacturer-identified nanotechnology-based consumer products currently on the market.[2] As from the start of 2011, the inventory holds more than a thousand entries in very diverse categories.

The very large and diverse array of applications as well as its enabling nature suggests that the term nanotechnology is more an abstraction than a clearly defined field of technology (Davis 2007). Nanotechnology is not so much an industry nor is it a basic technology in the classical sense with a clearly defined field. Nanotechnology is a collection of tools and approaches that can be adopted for specific applications. Nanotechnology is called an "enabling technology," since it can be applied to drive developments in derivative technologies in diverse fields.

Nevertheless, the term is widely used as a kind or shorthand representation of product and processes that utilize nanoscale properties. There is currently no widely accepted definition of nanotechnology (Balogh 2010). The lack of agreement on a definition that is shared by all stakeholders (including manufacturers, regulators, enforcement bodies, and consumers) has proved to be challenging because it forms a hurdle in developing policies and setting up proper regulations (Romig et al. 2007). In comparing the definitions proposed by various authors, it becomes clear that nanotechnology refers to at least three considerations:

- The dimension in the nanoscale range
- Properties or phenomena that can be attributed to this dimension
- Intentional exploitation of these properties or phenomena

[2]The online inventory can be found at http://www.nanotechproject.org/inventories/

Here we will use a working definition closely related to the broad definition provided by the Royal Society (Royal Society and Royal Academy of Engineering 2004) that entails these three common considerations. We define nanotechnology as design, production, and application of structures, devices, and systems by controlling shape and size with a least one critical dimension in the order of 1–100 nm. In this respect nanomaterials are intentionally engineered with at least one critical dimension in the order of 1–100 nm for a specific property. We refer to nanoparticles when we mean nanomaterials of specific shapes, such as dots, bars, dendrimers, colloids, tubes, and wires.

Nanomaterials possess properties different from their constitute materials of molecular or macroscopic size, because several physical phenomena become more pronounced at the nanoscale. These pronounced properties can be the result of quantum effects that play a more dominant role in the nanosize range compared to larger objects or they can result from the highly different physical properties, such as increased surface area per unit of substance compared to macroscopic systems. For example, titanium dioxide powder is known for its white appearance, while nanosized titanium dioxide is transparent. Furthermore, it should be noted that the 1–100 nm size range is in the order of magnitude at which many biological systems operate. These properties of nanomaterials enable applications, which are not possible using molecular or macroscopically sized materials. To reach the nanolevel there are two basic approaches in nanotechnology. In the "bottom-up" approach materials and devices are constructed from molecular components, essentially by building nanomaterials atom by atom. For this approach molecular self-assembly is very important. The "top-down" approach is the refinement of techniques and practices to the point that they reach the nanolevel and in essence the nanomaterial is constructed by breaking down larger objects.

## Short History of Nanotechnology

Nanotechnology is a relatively recent development and its roots are frequently associated with the presentation that famous physicist Richard Feynman gave at Caltech in 1959 entitled "There's Plenty of Room at the Bottom" (Feynman 1960). Even though Feynman did not use the term nanotechnology and his talk did not receive much attention until the beginning of the 1990s (Toumey 2009), it is considered inspirational to the field of nanotechnology. In fact, it was Norio Taniguchi of Tokyo University of Science who first coined the term "nanotechnology" at a conference in 1974 (Taniguchi 1974). The term got popularized by Kim Eric Drexler in his book *Engines of Creation: The Coming Era of Nanotechnology* published in 1986 and got well known in the scientific community once the journal *Nanotechnology* was founded in 1989.

The most well-known nanomaterials are fullerenes, such as the buckyballs and carbon nanotubes. Sir Harold Walter Kroto, Richard Errett Smalley, and Robert Floyd Curl, who share the Nobel Prize in Chemistry for this breakthrough, discovered buckminsterfullerene in 1985. The discovery of carbon nanotubes is attributed

to Sumio Iijima in 1991, although Roger Bacon at Union Carbide and Russian scientists behind the Iron Curtain were already working on such carbon fibers in the 1950s and 1960s (Colbert and Smalley 2002). From a more technical perspective, the field of nanotechnology started to develop in the 1980s with the invention of the scanning tunneling microscope and the atomic force microscope. The advances in microscopy are vividly illustrated by the Don Eigler and Erhard Schweizer paper in *Nature* of 1990 that reported that they had spelled out the name "IBM" with 35 xenon atoms.

The event that got the field off the ground was the huge-scale National Nanotechnology Initiative (NNI) project of the United States in 2000. The US commitment to nanotechnological development is significant with the cumulative governmental funding up to 2010 in the order of 12 billion US dollar, which makes it only rivaled by the NASA space program. The market size of nanotechnology-enabled products is estimated at about 250 billion US dollars worldwide. Development analysis projects that the number of nanotechnology products will achieve a 3 trillion US dollar market and 6 million workers by 2020 (Roco et al. 2010).

Together with the first conception of nanotechnology in the mid-1980s, there was mention of the possible ethical, legal, and social implications (ELSI). When large-scale organizations emerged to promote research and development of nanotechnology in the late 1990s – such as the Foresight Institute, the US National Nanotechnology Initiative, and the EU nanotechnology program – funding of accompanying research in ELSI as well as environmental, health, and safety (EHS) of nanotechnology became the norm. The first major attempt to evaluate the social and ethical implications of the nanotechnology development was a workshop of the National Science Foundation in 2000. The most influential report on the possible implications of nanotechnology was put forward by the Royal Society and Royal Academy of Engineering (2004). The possible negative effects of nanotechnology were popularized by many end-of-the-world scenarios, for example, the gray goo of out-of-control self-replicating robots that consume all matter on the Earth in the novel *Engines of Creation* by Drexler or the swarm of sentient nanorobots in the novel *Prey* by Michael Crichton.

## Central Moral Values and Value Issues

As indicated in the introduction, most of the moral values and related moral issues at stake with nanotechnology are not fundamentally new nor are they unique to nanotechnology. For example, Kuiken (2011) has argued that "[t]he ethical issues surrounding nanomedicine [. . .] are not new, but rather the techniques and science to achieve these improvements are new." This is not to say that the concerns raised by these moral issues can be dismissed as "nothing new." Novelty of a moral issue in general seems to be a poor guide for allocation of ethical inquiry. We would rather argue that although the novel moral issues seem philosophically more interesting, the nonunique moral issues also deserve attention. Since

nanotechnology is an enabling technology, it can intensify these existing nonunique moral issues or provide ways to address these issues. Furthermore, the application of nanotechnology could result into situations in which moral values are combined in new ways, come into conflict in unprecedented manners, or require a reconsideration of the perception of the moral value at stake, due to the altered context of the situation brought about by nanotechnology.

Nanotechnology is an enabling technology and therefore whole arrays of moral values are at stake. The moral issues arise from the integration of nanotechnology with the socio-technical context in which it is emerging. Hence, the nanomaterial by itself does not have an obvious recognizable connection with application and can only be used in a limited way to identify value issues. A more promising route is to address the moral values from the perspective of nanotechnological applications. With a perspective on applications, it is more straightforward to investigate relevant impacts and therefore reflect on the value issues at stake. In other words, the values, which are at stake in nanotechnology, are dependent on the context of its application. For example, carbon nanotubes are being utilized in displays, probes for atomic force microscopes, sensors, as well as lightweight composites for bikes, boats, windmills, and space travel. All these applications give raise to different moral issues with specific emphasis on particular moral values. In accordance, we will thus differentiate the moral values with regard to the specific applications.

To provide further structure to our analysis, we will distinguish between short-term and long-term applications of nanotechnology. With short-term applications we mean the applications of nanotechnology, which are currently on the market or have high promise to reach market in the near future. Examples of current applications of nanotechnology are silver nanoparticles as anti-odor agent in textiles and titanium dioxide nanoparticles as UV filters in sunscreens. In contrast, long-term applications are envisioned utilizations of nanotechnology in the far future. In the short-term an important role will be played by moral values such as equity, justice, privacy, responsibility, safety, and sustainability, while in the long-term the focus will be on other values such as human dignity, integrity of human nature, and intergenerational justice.

Our analysis does not address moral issues that can arise during the process of research related to nanoscience. Examples are safety issues with regard to the use of nanoparticles within the laboratory and accountability issues with authorship of publications. The focus is on the moral issues of the applications of nanotechnology in the context of the product life cycle as well as the way designers, engineers, and developers are able to shape the nanotechnology-enabled product with respect to the moral issues at stake during its life cycle.

## Values in the Short Term

Various authors have already investigated the moral values that play a central role in applications of nanotechnology (Choi 2003; Lewenstein 2005; Malsch and Hvidtfelt-Nielsen 2010; Sandler 2009; Royal Society and Royal Academy of Engineering 2004). The most frequently mentioned moral values associated with nanotechnology provided by these authors are accountability, animal welfare,

autonomy, fairness, equity, justice, nonmaleficence, privacy, quality of life, responsibility, safety, security, sustainability, transparency, and user friendliness.

Of these moral values, accountability, fairness, equity, justice, nonmaleficence, and responsibility are related to the power distribution and social interactions that shape the coexistence of technology and society. Since nanotechnology is an enabling technology, the socio-technical issues related to these values are legion and span a very wide range. The issues include lack of accountability in industrial as well as military research, unequal access to specific health-care treatments, and externalization of environmental costs of manufacturing methods (Sandler 2009). Nanotechnology is not the cause of these problems, in the sense that it is not the cause of the socio-technical issue, because the issue was inherent in the technology that is enabled by nanotechnology as well as the technology's social embedding. Nevertheless, the introduction of nanotechnology in the socio-technical context can intensify the existing problems due to the distinctive properties and functionalities that nanotechnology can provide. The flip side is that these features of nanotechnology can also provide opportunities to contribute in addressing the socio-technical issues. For example, currently there is an uneven utilization of technology at the international level, which leads to issues of equity. As nanotechnology enables existing technologies, it seems likely that countries having a high utilization of technology will benefit the most of the development of nanotechnology, which would lead to an exacerbation of the inequalities. This concern has been termed the "nano-divide" and concerns have been raised about further uneven power and wealth distribution.

The other moral values, which are not directly related to the above described socio-technical issues, such as animal welfare, autonomy, privacy, quality of life, safety, security, sustainability, transparency, and user friendliness are highly dependent on the specific application that nanotechnology enables. Table 1 gives an impression of the sort of moral values that are at stake here. This table is based on an extensive literature study of which the most important references can be found at the end of this chapter. No effort was made to make a systematic inventory; Table 1 shows the variety of values only, not a precise distribution of values over topics. For example, privacy is a key value at stake in ICT applications using nanotechnology for storing personal information, while it is of a very limited importance with deodorants that utilize nanomaterials as active ingredient.

## Values in the Long Term

Ethical inquiries into the long-term developments of nanotechnology commonly revolve around the manipulation of individual atoms and molecules that would lead to the ability to build any desired construction, ranging from nanoartifacts at the nanoscale to artifacts at micro- and macro-level. The one-by-one atom construction of larger artifacts would, of course in theory, require a very long time, as billions of atoms need to be placed in position. To solve this problem, the idea of general assemblers has been developed. These assemblers are in concept very similar to ribosomes in nature. They serve as machines that first multiply themselves and their exponentially growing "offspring" builds the artifact. An animation called nanofactory was published on YouTube to illustrate how a laptop computer could

**Table 1** Selection of short-term application of nanotechnology with their most prominent moral value(s) at stake in the current debate

| Technological sector | Application | Key moral value |
|---|---|---|
| Agricultural | Cattle monitoring | Animal welfare |
| | Product identification tags | Security, privacy |
| | Nutrient delivery | Safety |
| | Shelf-life-enhancing packaging | Transparency, safety |
| Chemical industry | Reaction catalysis | Sustainability |
| Construction | Barnacle-resistant coatings | Sustainability |
| | Self-cleaning surfaces | User friendliness |
| | Weather-resistant adhesives | Sustainability |
| Cosmetics | Anti-odor creams and sprays | Safety |
| | UV filter for sunscreens | Transparency, safety |
| Energy | Foldable solar cells | Sustainability |
| | Improved energy storage | Sustainability |
| Health care | Antimicrobial agent | Safety |
| | Diagnostic sensors | Privacy, safety |
| | Drug delivery | Safety, quality of life |
| | Surgical implants | Autonomy, quality of life |
| Information technology | Energy-efficient displays | Sustainability |
| | Information storage | Privacy, security |
| Textiles | Anti-odor | Safety |
| | Chemical protection | Security |
| | Water resistance | User friendliness |
| Transport | Fuel additive to increase efficiency | Sustainability |
| | Lightweight materials | Sustainability |

be built that way.[3] This development is still very speculative; nevertheless in the ethical debate, it is assumed that it makes sense to reflect on this development, because if it would be realized, it would have great consequences and many moral values would be at stake.

The primary domain of ethical concern seems to be that of medical technologies. The most far-reaching expectations of long-term nanotechnology developments are that it will be possible to repair human tissue so that life can be prolonged almost at will. This would have a great impact on human beings, as now one of its perhaps most important characteristics is its mortality.[4] Transhumanists welcome this development, but the question can be raised if humans will be able to make sense of life if it lasts for maybe hundreds of years. This permanent change in human

[3]The animation can be viewed at http://www.youtube.com/watch?v=zqyZ9bFl_qg and was sponsored by Nanorex, Inc.

[4]In the science fiction movie *Bicentennial Man*, this is even mentioned as the ultimate distinction between robots and humans. For a reflection on the way science fiction movies deal with the theme of blurring boundaries between humans and machines, see Cornea (2008).

potential is an example of what is called "human enhancement" (Lin and Allhoff 2008). Rather than restoring health in a situation of illness, human enhancement aims at enhancing human capabilities, both physical and mental. An issue at stake here is the possibility of a social divide: those who can afford to be enhanced may get control over others.[5]

Another development that would have a great impact on the nature of human existence is the possibility of making direct connections between the human brain and a computer. It is already possible to make a direct connection between nerve cells and devices for seeing/hearing and even an electrical wheelchair. Nonetheless, connecting the brain to a computer and thus being able to "read" what is in our mind would raise ethical question about the integrity of our human existence. Furthermore, the ability not only to manipulate the human body but also to have detailed knowledge about its state by means of complete DNA analyses using lab-on-chip devices could have as a consequence that we will be judged by our DNA. Already now, we see objections when insurance companies use medical data to determine the insurance rates one has to pay. Many would probably see being judged by one's DNA as a degradation of human dignity.[6]

The possibility of a new asbestos problem that was already mentioned in the previous section becomes more pressing when the long-term development of nanotechnology would lead to the possibility of creating extremely small devices that can invade the human body, e.g., in the veins to open obstructed arteries. If complete control of such devices is not guaranteed, they may get lost in the body and cause unpredictable damage. The same holds for nanodrugs that have a special coating that dissolves only at places where there are certain chemical substances that indicate the presence of a diseased cell. What will happen to the coating once it has dissolved? Do we know for sure it will not harm? Here, the value of safety is at stake.

## Approaches in Designing for Values

As in the previous section, we will make a distinction between short term and long term. For the short-term "Design for Values" approaches, we focus on available approaches which deal with designs of nanotechnological utilizations that have high promise to reach market in the near future, giving special emphasis toward the moral values identified in section "Values in the Short Term". For the long term we look at approaches that cope with envisioned applications of nanotechnology in the distant future.

[5]This is not a new concern. It was expressed, for instance, already by C.S. Lewis in his book *The Abolition of Man*. At that time he was referring to the use of eugenics by the Nazis, but his objections seem strikingly applicable to human enhancement as he explicitly writes about the creation of humans with enhanced capabilities.

[6]Here, again, we see science fiction movies playing with that theme, for instance, the movie *Gattaca* in which a man can only participate in space travel if he delivers a friend's blood, hair, skin cell, and urine samples because he himself has a defect in his DNA.

## Short-Term Approaches

Nanotechnology is one of the first technological developments in which funding agencies – like the National Nanotechnology Initiative (NNI) in the USA, the Framework Programs of the EU, and NanoNed in the Netherlands – required accompanying ethical (ELSI, ethical, legal, and social issues, and EHS, environment, health, and safety) research. Most of these efforts are directed at specific parts of EHS research, such as nanotoxicity, mobility of nanoscale materials, and workplace practice. In ELSI the focus is mainly on regulatory capacity, outreach, and public acceptance. Other efforts in ELSI research that accompanies nanotechnological R&D that involve moral issues are mainly aimed at the engagement of the public with developments in nanotechnology. So these efforts primarily focus on communicating with the general public and involving public opinion in policy setting. Hence, they can offer a forum for debate on ethical issues of nanotechnology, though they do not directly strive to develop approaches to Design for Values.

Overall, it is not an overstatement to say that within the ELSI research into responsible development of nanotechnology, the perspective of design has received little attention. Approaches to Design for Values that are specific to nanotechnology are missing, due to this limited scholarly effort into this field. It should be noted that the current funding focus on ELSI research aimed at engagement studies is not so surprising after the backlash in the field of biotechnology with genetic modification and the general association of ethics in relation to technology with prohibitions and restraints. This association is most commonly expressed in the sense that ethical issues should be addressed to prevent negative effects on the development and implementation of the technology. In essence, a proscriptive role [7] is assigned to ethical inquiry. However, we would like to stress that moral values can also be used in a positive sense. In other words, moral values can be used to encourage and guide the "good" development of technology, which requires one to identifying what is desirable and worth of pursuing as individual and for society.

As described in section "Central Moral Values and Value Issues" there is a whole range of moral values at stake in the application of nanotechnology. However, only a few authors have described approaches in which these values could be used in a positive sense for the design of products utilizing nanomaterials. In the following sections we will describe three initial attempts to Design for Values tailored to the field of nanotechnology. Firstly, we will describe the "safety by design" approach described by Christopher M. Kelty. Next, the attempt of Catherine J. Murphy is discussed, who puts forward sustainability as a design criterion for the production and usage of nanoparticles. Finally, we will explain the closely related approach of Johannes F. Jacobs et al. in which green chemistry principles are transferred to nanotechnological design practice.

[7] Here, we use the distinction between prescriptive and proscriptive morality. Proscriptive morality is focused on what we ought not to do and is inhibition based, while prescriptive is focused on what we ought to do and is activation based.

### Safety by Design

Kelty (2009) describes a "safety by design" approach based on an ethnographic study on work done by the National Science Foundation Center for Biological and Environmental Nanotechnology (CBEN) at Rice University in Houston, Texas, and the International Council on Nanotechnology (ICON) on the toxicity of buckminsterfullerenes. The ICON established the idea behind the approach and it was further developed together with the CBEN. The approach is an "attempt to make 'safety' a fundamental property of new nanomaterials: 'safety by design'" (Kelty 2009, p. 81) and it is attributed to the work of Vicki Colvin on the $C_{60}$ buckyball. In essence, the described method is a way to go beyond the toxicity implications after the fact of production and to design by identifying engineerable properties of new material with respect to toxicity.

In the "safety by design" approach, safety must be a property of nanomaterials of equal value to other "fundamental" physical and chemical properties, like specific gravity, thermal conductivity, magnetic permeability, and solubility in water. Safety is thus defined similarly to fundamental terms by bringing in concepts from biology and environmental sciences. In doing so, the safety can be tuned and controlled just like the physical properties of the material product.

Making safety a property on par with other accepted physical and chemical properties is a radial break away from the traditional conception of safety. For toxicologists, safety is a spectrum of risks resulting in adverse effects for living organisms; the risk spectrum concerns man-made materials in relation to complex ecosystems for environmental scientist, while for process engineers safety is inherent to the type and conditions of the manufacturing process as well as the disposal of waste. It is also a breakaway from the general idea that one first develops a beneficial application, before testing and verification of potential negative consequences. This idea is most prominent in the notion that it is the responsibility of regularity agencies and corporations to test and judge the safety of nanomaterials before commercialization, not the responsibility of scientists that discover and characterize these nanomaterials.

For the safety by design approach to work, it requires that toxicity must not solely be placed in a "language of hazard, exposure, and risk" but also in a "language of engineering and control of matter." In other words, the toxicity of nanomaterials "exists, but it is an interesting problem for materials chemists and nanotechnologists – one related to the properties of the material, its derivatizations, and its surface chemistry" (Kelty 2009). In light of the "safety by design" approach, the research into the toxicity of nanomaterials is one of concern ("is the material toxic?") and control ("how can the toxicity be modified?"). The approach thus implies that while toxicological research is essential for discerning how to engineer toward safety, it is insufficient to only inquire about the risks and hazards of every new material. The approach thus reopens inquiries about the predictability of toxicological effects; however, to date very little data exists to effectively implement the approach directly in engineering design. Nonetheless, we think this approach can be a fruitful starting point for research and development to incorporate the value of safety as a driver.

We think that the approach has a lot in common with the "inherent safety" concept that is mainly used in process industry to make an inherent safer design and would like to refer the reader to the chapter on safety by Neelke Doorn and Sven Ove Hansson in this volume (see "▶ Design for the Value of Safety". Nevertheless, it should be noted that nanotechnology opens the possibility to change the properties by designing the nanomaterial, while in other fields the focus is mainly on exchanging hazardous substances and processes for less harmful alternatives.

## Design for Sustainability

Catherine J. Murphy (2008) proposes that sustainability should be used as a design criterion for the synthesis as well as application of nanoparticles. She provides the example of quantum dot synthesis. Quantum dots are nanosized semiconductors that have interesting properties for lasers, light-emitting diodes, photodetectors, photo-imaging, solar cells, and transistors, due to confinement effects that result from their limited size. Most quantum dots are made of binary alloys such as cadmium selenide, cadmium sulfide, or cadmium telluride. However, the synthesis methods are far from sustainable. The feedstock used for the regular synthesis route is dimethylcadmium, which has several problems from a sustainability perspective, such as (a) the substance is very toxic, (b) is a known human carcinogen, and (c) poses explosion danger at temperatures used in the synthesis. Murphy shows that using sustainability as a design criterion can result in the discovery of more benign feedstock such as cadmium oxide or cadmium. She also puts forward investigations in manganese-doped zinc selenide as an alternative to the cadmium-based quantum dots, in an attempt to open up the design space for more sustainable production methods.

Murphy provides a second example with gold nanoparticles that have interesting optical properties that could be utilized in imaging technologies or as a chemical catalyst. Currently, these nanoparticles are produced using benzene and diborane, which are known to be toxic. Furthermore, the downstream processing requires huge amounts of organic solvents. Murphy (2008) shows that research with sustainability in mind generated a production process for these gold nanoparticles that replaced the two toxic substances with more benign alternatives, used less organic solvent for the membrane filtration, and decreased the overall production cost with a factor of about 100. Furthermore, she described ongoing research efforts with the aim to develop more sustainable processes for gold nanoparticle production that use water as the solvent, take place at room temperature, and utilize mild reducing agents by using surface for the particle growth.

As a general approach for the more sustainable production of metal nanoparticles, Murphy (2008) proposes the use of metal salts in a water solvent with biological reduction agents. These processes are in general more benign substances and mild operation conditions, in effect reducing energy usage and lowering the potential impact on workers as well as the environment. A second approach put forward by Murphy is coating the nanoparticles in such a way that they become more benign. This approach depends on the observation that most biological interaction at the nanoscale is highly dependent on the surface of the

nanoparticle instead of the composition of the core. Nonetheless, we find this second approach failing in two respects. First of all, the coating of nanoparticle makes recycling of the particles more difficult, because it is a mixture of substances. Secondly, the coating only provides a layer of protection that will inevitably fail over time instead of designing the particle to be inherently less poisonous.

Rightfully, Murphy also points toward the potential of nanomaterials in environmental friendly applications – such as an environmental remediation and solar cells – as a way toward the adoption of the sustainability criterion for the utilization of nanotechnology instead of only the production of nanomaterials. Nonetheless, we think that further research is necessary that incorporates the whole life cycle (including the production and disposal of the utilized nanomaterials) to see if such applications are overall more sustainable.

### Green Nanoprinciples

Like the design for sustainability approach by Catherine J. Murphy discussed above, some authors have taken inspiration from green chemistry, especially because in the recent years, "green chemistry" has been successfully utilized to reduce or eliminate the usage and generation of hazardous substances in the design, manufacture, and application of chemical products. For example, Lallie C. McKenzie and James E. Hutchison (2004) see an opportunity for the cross-fertilization between the fields of green chemistry and nanoscience. They state that "the principles of green chemistry can guide responsible development of nanoscience, while the new strategies of nanoscience can fuel the development of greener products and processes." The idea has inspired the term "green nanotechnology" to which topic a journal, named the *International Journal of Green Nanotechnology*, is dedicated since 2009.

Green chemistry is a set of 12 principles,[8] developed by Paul Anastas and John C. Warner (Anastas and Warner 1998), which can be used to guide engineering design in chemical technology toward safety and sustainability. To transfer the approach from chemical technology to nanotechnology, an abstraction is needed to translate the approach from one discipline to the other. Jacobs et al. (2010) propose to abstract the 12 principles of green chemistry into four general concepts, knowingly:

- Product safety
- Low environmental impact
- Material and energy efficiency
- Process safety

[8]These principles are (1) waste prevention, (2) atom economy, (3) less hazardous synthesis, (4) design for safer materials, (5) safer auxiliaries and solvents, (6) design for energy efficiency, (7) renewable resources, (8) reduce derivatives, (9) catalysis, (10) design for end of useful life, (11) real-time monitoring, and (12) inherent safer processes.

The concept of "product safety" entails the aim of designing nanoproducts in such a way that they represent a low potential for generating hazards while maintaining their desired function. The "safety by design" approach, as described by Christopher M. Kelty (see section "Safety by Design"), fits nicely with the safety value of this concept. The "low environmental impact" concept aims for a product design that incorporates a whole life cycle view. In other words, the concept looks for nanoproducts, which are produced from renewable resources and are reusable, recyclable, or degradable into non-environmentally persistent components. The third concept indicates a need for the conservation of utilized resources in as far as possible. The concept aims for the value of sustainability by maximizing the incorporation of material into the final product and minimizing the utilization of energy. The "process safety" principle aims at the value of safety from the perspective of the production process. The nanoproduct manufacturing process should inherently pose as little hazards as possible for the workers and environment as well as have adequate safety features lowering the risk of potential process hazards.

The approach of using existing knowledge and know-how of more established fields of technology in order to aim for the incorporation of moral values such as safety and sustainability into design of nanotechnology seems to be a fruitful way to prevent the reoccurrence of known moral issues with technological development.

## Long-Term Approaches

As stated in the introduction, for long-term developments a Design for Values approach is more difficult than for short-term developments because there is still speculation about what the artifacts to be designed will be like. Nevertheless, the terms "design" and even "design considerations" do feature in nanotechnology literature.[9] Ethical considerations are not yet found in such references, though. But the values at stake do seem to be clear (see section "Values in the Long Term"). The real challenge is to deal with the issue of traditional categories (natural-artificial, healthy-ill, human-machine, and the like) for ascribing values becoming problematic. Martijntje Smits has suggested using a strategy that she called "taming the monster." Here, the term "monster" refers to the fear people get when they come across products that cannot be immediately put into a certain traditional category (Smits 2006). This means that we have to redefine our categories such that the new technology can be characterized and understood in terms of the new categories. Although at first sight this seems an attractive option to deal with these problems, one can question if it does justice to the concerns one may have. Does redefining the categories solve the problem or does it walk away from them by means of a conceptual "trick"? Are these categories purely epistemic and is there really no ontic aspect to these categories? In other words, is the problem only in our thinking, or is it also in the reality outside our minds? (Table 2)

[9]For example, Merkle (1996), and Choi et al. (2010).

**Table 2** Challenged traditional categories of long-term application of nanotechnology

| Type 1 | Type 2 | Nature of confusion |
|---|---|---|
| Human | Machine | Extreme close connection between human and machine ("cyborg") |
| Natural | Artificial | Engineered processes that mirror exactly the natural processes |
| Healthy | Ill | State of knowing the chances of certain potential diseases becoming actual |
| Living | Nonliving | Building up tissue from scratch with unclear transition from nonliving to living |
| Mortal | Immortal | Extending the life span at will |

This table is based on Boenink et al. (2010)

Another difficulty for ethical reflection on long-term developments in nanotechnology was the difficulty to imagine possible effects. Here, too, a proposal has been done to solve this difficulty, namely, that of "techno-moral" scenarios (Boenink et al. 2010). This tool is meant to enhance imagination in cases where consequences of technology are not obvious. Of course, this tool functions primarily in the context of a consequentialist approach to ethical problems, and if one does not adhere to such an approach, the value may be limited. Both the "monster taming" and the "techno-moral scenario" approaches have the disadvantage that they only support the long-term development assessments, but they do not provide clues for Design for Values. At best, they help to gain insight into what values are at stake. As long as there values are ones that we know from the past or current ethical debates, the stage of "monster taming" and/or "techno-moral scenario" building can be followed by a stage in which existing approaches for Design for Values are applied, as then we are again in a known domain.

## Comparison and Evaluation

When we compare short- and long-term developments, we see that in the short term Design for Values plays a role in the nanotechnological developments, be it a relatively small one. In the long-term developments of nanotechnology, there is no concrete elaboration of the notion of Design for Values yet, but there are efforts to get more view on what values are at stake. Due to blurring of boundaries between traditional categories, it is difficult to relate values to categories as a preliminary step toward Design for Values. The extent to which category boundaries really will get blurred is, however, unclear as it is difficult to picture a realistic image of what the effects of nanotechnological developments might look like. However, scenario techniques, such as the techno-moral scenarios, may help to get more clarity here, and this may lead to taking the next step toward Design for Values, as the relation between values and (new) categories can then be identified.

## Experiences and Examples

As in the previous section, we make a similar distinction in time frame. The short term will be illustrated with the application on nanoparticles in sunscreens, while cyborgs will be the example of long-term nanotechnological developments.

### Nanoparticles for Sunscreens

Nanoparticles of titanium dioxide ($TiO_2$) are currently utilized in a wide variety of products. These $TiO_2$ particles are, for example, used as UV protective agents in cosmetic sunscreen and plastics but also as photocatalysts for the photodegradation of pollutants in wastewater and cancer treatments or as coating for "self-cleaning" windows. For this case study, we will focus on the sunscreen application because sunscreens containing nanosized $TiO_2$ are sold worldwide for over a decade now and it is one of the most widely known first-generation nanotechnological products.

As we are dealing with a cosmetic product, it is clear that the value of safety is at stake. Safety is here mostly related with possible negative effects on human health but also to the hazards associated with the manufacturing process. When considering the whole life cycle of the product, it is obvious that sustainability is also a moral value that is at stake with the manufacturing process, required resources, and disposal. Jacobs et al. (2010) have shown that by using the "green nanoprinciples" for the current production methods as well as for the design of the final product, some noteworthy advances can be made in designing for the moral values of safety (see section "Green Nanoprinciples"). The analysis shows that there is still a large room for improvement left with regard to safety and especially sustainability. For example, Jacobs et al. discuss the widely acknowledged problem with the formation of reactive oxygen species (ROS) when $TiO_2$ nanoparticles are excited with UV light. These formed ROS are known to cause negative health effects on humans and pose ecological risks. The issue can be reduced by designing the nanoparticle in such a way that it consists of a crystal morphology that is less photoactive and hence produces less ROS. Besides, doping the particles with another metal or coating the $TiO_2$ surface with silica, alumina, and/or polymers can reduce the production of ROS. Most of these ways to reduce the ROS formation are currently employed by production companies for $TiO_2$ nanoparticles intended for sunscreen applications.

On the other hand, Jacobs et al. (2010) show that the current manufacturing practice does not follow a design for sustainability approach. One issue is that the raw materials for the production are obtained from nonrenewable resources, such as the mining of titanium containing ore for natural deposits. Other sustainability issues are the use of chlorine gas as well as extreme operational conditions posing environmental risks as well as a high consumption of energy in the form of combustion agents, such as ethane or hydrogen. It should be noted that the used high temperatures – in the range of 900 °C – also pose hazards to the workers.

Overall, it seems that although there are some examples of application of Design for Values with respect to safety for first-generation nanoparticle-containing products, only minor efforts for the design for sustainability have been undertaken. Other moral values that are potentially at stake have received even less attention, not in the least because there is currently a clear lack of Design for Values approaches specific to nanotechnology.

## Cyborgs

One of the promises of the application of nanotechnologies in the domain of health care is the enhancement of human capabilities through extremely smooth transitions from human beings to artifacts. Human brain cells may be directly connected to computer wires. This will create a hybrid being that most commonly is called a cyborg. Transhumanists hope that this will also enable us to store our mind in hardware so that we can live on forever. Ray Kurzweil in this context uses the term "singularity," the complete integration of humans and machines (Kurzweil 2005). Ethical questions have been raised about this and some suggestions have been made about Design for Values considerations. Although the term eugenics is carefully avoided in most writings about human enhancement, no doubt because of its negative connotations, a fear for the development of a sort of super-being is sometimes expressed. In itself the idea of human enhancement through technology is far from new. The philosopher Ernst Kapp already suggested that all technology in some way or another is an extension of the human body.[10] Also the idea of extending the human mind through technology has been suggested, for example, in the extended mind theory developed by Andrew and David Chalmers. But in those writings, all examples are such that it is well possible to indicate where the human part of the human-machine combination ends and where the machine part begins. This, however, would be much more problematic in the case of cyborgs and the singularity. This causes category boundary definition problems, as discussed in section "Long-Term Approaches", particularly in the human-machine and mortal-immortal categories.

One of the primary values at stake here is human dignity (Rubin 2008). Some authors have suggested design criteria for human-machine combinations of a cyborg-like nature that aim at preserving this dignity. Jeff Wildgen, for instance, refers to Asimov's "classic" three laws[11] for robot design as a possible set of criteria that also hold for singularity-related designs (Wildgen 2011). Machiel van der Loos (2007) also refers to Asimov's laws and suggests that cyborgs will be designed to

[10]See the resent analysis by Lawson (2010).

[11]These laws are as follows: (1) a robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) a robot must obey orders given to it by human beings, except where such orders would conflict with the first law; and (3) a robot must protect its own existence as long as such protection does not conflict with the first or second law. Asimov introduced these laws in a 1942 short story called *Runaround*.

have agency, and for that reason ethical constraints should be in the list of requirements, just like Asimov suggested for robots. He mentions the condition of the cyborg having control over the implants as another dignity-related ethical requirement for cyborg design. This also relates to the integrity of the human personality as a moral value at stake here. According to Kevin Warwick – who had a silicon chip transponder implanted in his upper left arm himself – merging human and machine will have an impact on the individual's consciousness and personality (Warwick 2003). The option of linking persons through the transponders, for instance, means that they are no longer individuals but very intimately connected to other people's minds. Warwick suggests that cyborgs may develop their own type of consciousness and their own morality related to that.

## Open Issues and Further Work

Research initiatives on nanotechnology can be found all over the world. Even upcoming economies such as Argentina, Brazil, China, India, the Philippines, South Africa, and Thailand are now investing in nanoscience and technology (Salamanca-Buentello et al. 2005). Nanotechnology is turning global and the cultural diversity of perceptions of ethical issues due to differences in cultural heritage, economic conditions, as well as political situations should thus also be addressed (Schummer 2006). Currently, the majority of scholars working on Design for Values specifically for nanotechnology are based in the USA and Europe. Although the presented approaches are broad enough to embrace some cultural diversity, there is a need for Design for Values approaches from a non-Western perspective.

As nanotechnology is a relatively new technological field, its development is still plagued by uncertainties. These uncertainties are the result of lack of knowledge, ignorance, and complexity. Ignorance, also called the "unknown unknown," is a very troubling part of uncertainty of a novel technology because we do not know what we have to prepare ourselves for. A Design for Values approach should be able to deal with these kinds of uncertainty that plague the conception and initial implementation of a technology. Vermaas et al. (2011) have suggested that the designers should take into account robustness, flexibility, and transparency to deal with this issue. We think that adaptability over time, dependent on the new information that comes available, is an appropriate starting point for a Design for Values approach that wants to deal with this uncertainty issue. Alternatively one could choose to wait for further development of the technology before aiming at Design for Values approaches. However, the "Collingridge dilemma" (Collingridge 1980) makes clear that the impact of steering the development in light of moral values is the greatest in the initial phases of development, but unfortunately there is a limited amount of knowledge available at that moment.

A complicating issue with nanotechnology is the diversity of materials and techniques that it represents. Nanomaterials themselves can be the product of nanotechnology or could be used to manufacture products that do not contain nanomaterials. Even when only nanomaterials are considered, the diversity is

extremely large as a result of the numerous ways a nanoparticle of a given composition can be made functional for specific applications. A nanoparticle of a given composition can have various morphologies, crystal structures, size distributions, and agglomeration or aggregation states. This heterogeneity asks for a Design for Values approach that can deal with this diversity and can incorporate various analyses, which are made on a case-by-case basis. For example, to evaluate the toxicity risk of a chemical substance, it is needed to assess the toxicity hazard as well as the exposure of a nanoparticle. In current chemical risk assessment, the exposure is characterized with a measure of concentration; however, such a measure is not always adequate for nanoparticles due to the abovementioned issues of size distribution, shape, aggregation, etc. A design for safety approach thus should be flexible enough to incorporate this diversity.

For the long-term considerations, the issue of seemingly confused category boundaries needs more exploration. As Geertsema has pointed out, whether one accepts the blurring of category boundaries depends on one's ontological assumptions (Geertsema 2006). If this is the case, the problem of confused boundaries may exist only for certain ontological stances and not for others. This will have consequences, of course, for the moral questions related to these boundaries.

## Conclusion

In this chapter we have shown that nanotechnology is a field of new and emerging technology that brings about relatively new ethical issues, in particular for the long term. For the short term, no fundamentally novel values are at stake and there are some first initiatives aimed at Design for Values. With respect to the long term, ascribing values to categories is hampered by the fact that some traditional category boundaries are blurred in the case certain expectations appear to be realizable. In particular molecular nanotechnology may cause truly novel ethical issues due to the blurring of boundaries. Scenario techniques can be used to get a clearer picture of what the technology may look like and this may speed up the development of Design for Values.

## Cross-References

► Design for the Value of Safety
► Design for the Value of Sustainability

## References

Anastas P, Warner JC (1998) Green chemistry: theory and practice. Oxford University Press, New York

Balogh LP (2010) Why do we have so many definitions for nanoscience and nanotechnology? Nanomed Nanotechnol Biol Med 6(3):397–398

Boenink M, Swierstra T, Stemerdink D (2010) Anticipating the interaction between technology and morality: A scenario study of experimenting with humans in bionanotechnology. Stud Ethics Law Technol 4(2):1–38. Article 4
Choi K (2003) Ethical issues of nanotechnology development in the Asia-Pacific region. In: Bergstrom I (ed.) Ethics in Asia-Pacific. Regional Bureau for Education UNESCO, Bangkok, pp 327–376
Choi HS, Liu W, Nasr F, Misra K, Bawendi MG, Frangioni JV (2010) Design considerations for tumor-targeted nanoparticles. Nat Nanotechnol 5:42–47
Colbert D, Smalley R (2002) Past, present and future of fullerene nanotubes: Buckytubes. In: Osawa E (ed) Perspectives of fullerene nanotechnology. Kluwer, Dordrecht, pp 3–10
Collingridge D (1980) The social control of technology. Frances Printer, London
Committee on Technology (2000) National nanotechnology initiative: leading to the next industrial revolution. Interagency Working Group on Nanoscience, Engineering and Technology, Washington, DC
Cornea C (2008) Figurations of the cyborg in contemporary science fiction. In: Seed D (ed) A companion to science fiction. Blackwell, Oxford, pp 275–288
Davis JM (2007) How to assess the risks of nanotechnology: learning from past experience. J Nanosci Nanotechnol 7(2):402–409
de Vries MJ (2006) Analyzing the complexity of nanotechnology. In: Schummer J, Baird D (eds) Nanotechnology challenges: implications for philosophy, ethics and society. World Scientific, Singapore, pp 165–178
de Vries MJ (2008) A multi-disciplinary approach to technoethics. In: Luppicini R, Adell R (eds) Handbook of research on technoethics. Information Science Reference, Hersey, pp 20–31
Ferrari A (2010) Developments in the debate on nanoethics: traditional approaches and the need for new kinds of analysis. Nanoethics 4(1):27–52
Feynman RP (1960) There's plenty of room at the bottom: an invitation to enter a new field of physics. Eng Sci 23(5):22–36, Transcript of talk delivered at California Institute of Technology, Pasadena, 29 December 1959
Geertsema H (2006) Cyborg: myth or reality? Zygon 41(2):289–328
Grunwald A (2010) From speculative nanoethics to explorative philosophy of nanotechnology. Nanoethics 4(2):91–101
Jacobs JF, van de Poel I, Osseweijer P (2010) Towards safety and sustainability by design: nano-sized $TiO_2$ in sunscreens. In: Fiedeler U et al (eds) Understanding nanotechnology: philosophy, policy and publics. IOS Press, Amsterdam, pp 187–198
Kelty CM (2009) Beyond implications and applications: the story of 'safety by design'. Nanoethics 3(2):79–96
Kuiken T (2011) Nanomedicine and ethics: is there anything new or unique? Nanomed Nanobiotechnol 3(2):111–118
Kurzweil R (2005) The singularity is near. Penguin Books, New York
Lawson C (2010) Technology and the extension of human capabilities. J Theory Soc Behav 40(2):207–223
Lewenstein BV (2005) What counts as a "social and ethical issue" in nanotechnology? Hyle Int J Philos Chem 11(1):5–18
Lin P, Allhoff F (2008) Untangling the debate: the ethics of human enhancement. Nanoethics 2(3):251–264
Malanowski N et al (2006) Growth market nanotechnology: an analysis of technology and innovation. Wiley-VCH Verslag GmbH, Weinheim
Malsch I, Hvidtfelt-Nielsen K (2010) Nanobioethics 2nd annual report on ethical and societal aspects of nanotechnology. Report of ObservatoryNano, www.nanopinion.eu
McGinn RE (2010) What's different, ethically, about nanotechnology?: foundational questions and answers. Nanoethics 4(2):115–128
McKenzie LC, Hutchison JE (2004) Green nanoscience. Chemistry Today, Sept, pp 30–33
Merkle RC (1996) Design considerations for an assembler. Nanotechnology 7:210–215

Murphy CJ (2008) Sustainability as an emerging design criterion in nanoparticle synthesis and applications. J Mater Chem 18(19):2173–2176
Nordmann A (2007) If and then: a critique of speculative nanoethics. Nanoethics 1(1):31–46
Preston CJ et al (2010) The novelty of nano and the regulatory challenge of newness. Nanoethics 4(1):13–26
Rensselaer (2004) Nanotechnology sector report: technology roadmap project. In: Report by the Center for Economic Growth and the Lally School of Management and Technology, www.ceg.org
Roco MC et al (2010) Nanotechnology research directions for societal needs in 2020: retrospective and outlook. In: Science policy reports. Springer, Dordrecht
Romig AD et al (2007) An introduction to nanotechnology policy: opportunities and constraints for emerging and established economies. Technol Forecast Soc Change 74(9):1634–1642
Royal Society & Royal Academy of Engineering (2004) Nanoscience and nanotechnologies: opportunities and uncertainties. Report of the Royal Society & the Royal Academy of Engineering Working Group, Plymouth, UK
Rubin C (2008) Human dignity and the future of man. In: The President's Council on Bioethics (ed) Human dignity and bioethics: essays commissioned by the president's council on bioethics. US Government Printing Office, Washington, DC
Salamanca-Buentello F et al (2005) Nanotechnology and the developing world. PLoS Med 2(5): e97
Sandler R (2009) Nanotechnology: the social and ethical issues. PEN 16 report. Woodrow Wilson International Center for Scholars
Schummer J (2006) Cultural diversity in nanotechnology ethics. Interdiscip Sci Rev 31(3):217–230
Smits M (2006) Taming monsters: the cultural domestication of new technologies. Technol Soc 28(4):489–504
Swierstra T, van Est R, Boenink M (2009) Taking care of the symbolic order: how converging technologies challenge our concepts. Nanoethics 3(3):269–280
Taniguchi N (1974) On the basic concept of 'nano-technology'. Proc Int Conf Prod Eng Tokyo 2:18–23
Toumey C (2009) Plenty of room, plenty of history. Nat Nanotechnol 4(12):783–784
Van de Poel I (2008) How should we do nanoethics?: a network approach for discerning ethical issues in nanotechnology. Nanoethics 2(1):25–38
van den Hoven J, Manders-Huits N (2009) Value-sensitive design. In: Berg Olsen JK, Perdesen SA, Hendricks VF (eds) A companion to the philosophy of technology. Wiley-Blackwell, Chichester, pp 477–480
Van der Loos HFM (2007) Design and engineering ethics considerations for neurotechnologies. Camb Q Healthc Ethics 16(3):303–307
Verbeek P-P (2009) Ambient intelligence and persuasive technology: the blurring boundaries between human and technology. Nanoethics 3(3):231–242
Vermaas P et al (2011) A philosophy of technology: from technical artefacts to sociotechnical systems. Synth Lect Eng Technol Soc 6(1):1–134
Warwick K (2003) Cyborg morals, cyborg values, cyborg ethics. Ethics Inf Technol 5(3):131–137
Wildgen J (2011) Ethical considerations of the approaching technological singularity. In: CSE 5290 – artificial intelligence. http://www.mycigroup.com/Documents/Library/Singularity_CSE5290_Wildgen.pdf. Accessed 30 Sept 2011

# Design for Values in Nuclear Technology

Behnam Taebi and Jan Leen Kloosterman

## Contents



**Abstract**

Safety has always been an important criterion for designing nuclear reactors, but in addition to safety, there are at least four other values that play a key role, namely, security (i.e., sabotage and proliferation), sustainability (i.e., environmental impacts, energy resource availability), economic viability

B. Taebi (✉)
Department of Philosophy, Faculty of Technology, Policy and Management, TU Delft, Delft, The Netherlands

Belfer Center for Science and International Affairs, John F. Kennedy School of Government, Harvard University, Cambridge, MA, USA
e-mail: b.taebi@tudelft.nl

J.L. Kloosterman
Department of Radiation Science and Technology (RST), Faculty of Applied Sciences, TU Delft,
e-mail: j.l.kloosterman@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_30

(i.e., embarking on new technology and its continuation), as well as intergenerational justice (i.e., what we leave behind for future generations). This chapter reviews the evolution of generations of nuclear reactors (I, II, III, III, and IV) in terms of these values. We argue that the Best Achievable Nuclear Reactor would maximally satisfy all these criteria, but the safest reactor is not always the most sustainable one, while the reactor that best guarantees resource durability could easily compromise safety and security. Since we cannot meet all these criteria simultaneously, choices and trade-offs need to be made. We highlight these choices by discussing three promising future reactor types, namely, the high-temperature reactor pebble-bed module (HTR-PM), the molten salt-cooled reactor (MSR) and the gas-cooled fast reactor (GFR).

**Keywords**

Safety • Sustainability • Security • Economic viability • Intergenerational justice

## Introduction

In December 2011 Bill Gates announced that he plans to invest one billion dollars to jointly develop a new nuclear reactor with the company TerraPower. This reactor is designed to be less expensive than the current reactors; it must run on abundantly available natural uranium, it must generate little waste, and, perhaps most importantly, "all these new designs will be incredibly safe," Gates emphasized (BBC 2011).

Gates' reactor seems to be the ideal nuclear power solution as it enables us to enjoy the benefits of nuclear power without being troubled by any of its drawbacks. So this reactor is assumed to carry low accident risks, to not require any proliferation sensitive enrichment of uranium, and to produce only a small volume of high-level waste. These claims are all made by the manufacturer, who estimates this reactor to be available around 2030.

In designing nuclear reactors, several criteria have played an important role: the possibility and the probability of core failure or meltdown, the kind of fuel needed, the amount of energy produced, the volume and lifetime of the remaining waste after operation, and, last but not least, the possibility of using the reactor to manufacture one of the key ingredients of a nuclear bomb, namely, weapon-grade nuclear material. The latter is perhaps among the oldest issues in nuclear technology. The world's first nuclear reactor was built in the 1940s to show the feasibility of producing plutonium, which then could be extracted from the irradiated fuel. Although this was primarily intended for the energy generation purposes, plutonium was used shortly after its discovery in the Nagasaki bomb (Seaborg 1962). The dual use of nuclear technology, alternatively known as *proliferation*, has been a central issue since the beginning of the civil use of nuclear power in the 1950s and the 1960s up until the present.

The aforementioned criteria are referred to as *values*, since they reflect how we perceive "the good" or how we want the world to be (Scanlon 1998). Values are very important in the design of nuclear reactors. However, we cannot always

accomplish all the "goods" at the same time so we need to make choices and trade-offs between the *good* we find to be more important indicating why we find it more important. It has been argued that the impossibility of accomplishing several values at the same time – or simply value conflicts – has fueled innovation in engineering design (Van de Poel 2009; see also the chapter "▶ Conflicting Values in Design for Values" in this Volume). In designing nuclear technology, there are at least five main values that play a key role, namely, safety (i.e., public health impacts), security (i.e., sabotage and proliferation), sustainability (i.e., environmental impacts, energy resource availability), economic viability (i.e., embarking on new technology and its continuation), and intergenerational justice (what we leave behind for future generations).

These values should be in balance with each other since they cannot always be simultaneously accomplished. Different societal, ethical, or political considerations could bring one of these values to the forefront. It is particularly interesting to see how nuclear accidents have affected the perception of nuclear safety which, in turn, has determined the evolution of nuclear reactors. For instance, the development of substantially safer nuclear reactors started after the Three Mile Island accident in Pennsylvania in 1979. Since the Fukushima disaster in 2011, "safety" seems again to be the leading value in design.

The chapter is organized as follows: in section "Nuclear Technology," we will first introduce nuclear power technology and its key component, the nuclear reactor. Special attention will be devoted to the historical evolution of safety as a key value in nuclear technology design; we will also discuss other values that are relevant when designing civilian nuclear technology. Section "Design for Nuclear Values" focuses on the design of several new nuclear reactors and how a preference for different values has resulted in different nuclear reactor designs. Section "Open Issues and Future Work" presents the open issues for further academic endeavor. Conclusions are presented in section "Conclusions."

## Nuclear Technology

Radioactivity was discovered by the end of the nineteenth century. Yet, it had little practical relevance until 1938 when the first fission reaction (i.e., splitting the nuclei by neutrons) was discovered. Since a fission chain reaction releases more than one free neutron, a fission chain reaction could be made self-sustaining. This technology was used in the WWII for the development of nuclear weapons, but soon thereafter the same physical principles were applied for civil purposes. The first non-weapon application was for the propulsion of submarines in 1953. In 1956, world's first nuclear plant for electric production started operation at Calder Hall in the UK (Tester et al. 2005, Chap. 8). In this section, we briefly introduce nuclear power production. More specifically we will focus on nuclear reactors and the nuclear fuel cycle in which those reactors have a key role. We will focus on the evolution of safety in nuclear reactor design; other relevant values in nuclear technologies will also be introduced.

## Nuclear Reactor

The reactor is a key technological component for the production of nuclear energy. The evolution of nuclear reactors is often denoted in terms of "generations" including I, II, III, III, and IV. Each generation is developed with certain features as leading design criteria; Table 1 summarizes – among other things – the leading values behind each generation of reactors. The first generation of nuclear reactors was considered "proof of concept" for civil nuclear power, and they include the prototypes from the 1950s and 1960s. The only Gen I reactor still in operation is the Wylfa Nuclear Power Station in Wales. Gen II reactors are commercialized power plants that were designed to be *economical* and *reliable*; their operation started in the 1960s. The Gen III reactors are "state-of-the-art design improvements" in the areas of fuel technology, thermal efficiency, and safety systems (Goldberg and Rosner 2011, p. 6). The Gen III are designed with safety as leading design criterion. In Gens III and III, *passively safe* reactors have been introduced that would not require active control of the operator for safety; in the remainder of this section we will further discuss this issue. Finally, Gen IV reactors present revolutionary design changes. Unlike its previous generations, Gen IV reactors are one to four decades away and they are being designed to reconcile several design criteria, such as sustainability, waste management benefits, nonproliferation, and safety. In section "Design for Nuclear Values" we will show how different design criteria have led to drastically different designs for Gen IV reactors.

The majority of the world's 435 reactors still in operation today comprise Gen II reactors. These reactors use light water ($^1H_2O$) as a coolant and moderator which is why they are referred to as light water reactors (LWR). Of the LWRs, 75 % are pressurized water reactors (PWR), originally designed for ship propulsion. The remainder of LWRs are boiling water reactors (BWR) (Tester et al. 2005, p. 374).

### The Historical Evolution of Safety in Reactor Design

Historically, safety has been one of the important driving forces behind serious changes in reactor design philosophy. Major nuclear accidents seem to have particularly affected people's thinking about reactor safety. After the core meltdown accident in Three Mile Island in Pennsylvania in 1979, David Lilienthal called upon nuclear technologists to design safer nuclear reactors whose cores could not melt (Lilienthal 1980). This proposal was first only met with skepticism, but it did provoke a discussion on the philosophy of nuclear reactor safety (Weinberg and Spiewak 1984).

Before moving toward designing safer systems, the skeptics first proposed reassessing the probability of core damage in existing TMI-type nuclear reactors. A couple of years before the Three Mile Island accident, the Atomic Energy Commission (AEC) had initiated a new study to coherently assess the safety of nuclear reactors by mapping all the events that could possibly lead to an accident and then assigning probabilities to each single event. The study, officially known as the Reactor Safety Study, was better known as the Rasmussen Report (NRC 1975), and the proposed method was termed probabilistic risk assessment (PRA) (Keller and Modarres 2005). The Rasmussen Report found the core damage frequency of a

**Table 1** The evolution of generations of nuclear reactors

| Generation | II | III | III+ | III+ | IV | IV |
|---|---|---|---|---|---|---|
| Reactor type – acronym | PWR and BWR[a] | ABWR | AP1000 | HTR-PM | GFR | MSR |
| Estimated core damage frequency (CDF)[b] (per reactor year) | $10^{-4}$–$10^{-5}$[c] | $1.6 \times 10^{-7}$[d] | $4.2 \times 10^{-7}$[e] | $5 \times 10^{-7}$[f] | N.A.[g] | N.A.[g] |
| Type of change in design | Default design | Small and incremental compared to BWR | Medium and incremental compared to PWR | Radical | Medium to radical | Very radical. Change in reactor technology |
| Leading values in design | Safety | Safety | Safety | Safety and economic viability | Sustainability | Sustainability |

[a]Both PWR and BWR are generally referred to as the LWRs

[b]The phrase "core damage frequency" does not always refer to the same phenomenon since different studies employ different methodologies and adhere to different basic assumptions. The common understanding of the term is "damage of the core as a whole" but it could also refer to the damage of small parts of the core or to single fuel pins (Leurs and Wit 2003, p. 136)

[c]Different estimations have been given for different types of reactors. Rasmussen, for instance estimates $2.6 \times 10^{-5}$ and $4.6 \times 10^{-5}$ respectively for a PWR and BWR (NRC 1975). Other reports provide slightly different estimations, such as $5 \times 10^{-5}$ (EC-DGXII 1994). Nowadays generation II reactors can have CDF values up to a factor of 5–10 lower due to technical updates

[d]This is the estimation of one of the designers/manufacturers of ABWR, General Electric Hitachi Nuclear Energy (GEH). The NRC confirms this estimation, but it refers to this estimation as the "core damage frequency for internal events"; the NRC study further distinguishes between CDF from internal floods (i.e., $7 \times 10^{-9}$) and for fire (i.e., $1 \times 10^{-6}$; NRC 1994)

[e]The manufacturer calculated this probability to be $4.2 \times 10^{-7}$. The large release frequency after a severe accident has been estimated to be $3.7 \times 10^{-8}$ (Schulz 2006, p. 1553)

[f]This is the probability of radionuclides being released rather than the probability of a meltdown occurring; meltdown is in principle ruled out in this design. This calculation refers to the PBMR, but since the design characteristics with an HTR-PM are not substantially different, the estimation is probably a good indication of the probability of radionuclide release (Silady et al. 1991, p. 421)

LWR to be approximately $5 \times 10^{-5}$ per reactor year. An analysis of the actual precursors to potentially serious events in operating reactors between 1969 and 1979 suggested, however, a more pessimistic probability, namely, $10^{-3}$ per reactor year (Minarick and Kukielka 1982). Taking these semiempirical results into account, Spiewak and Weinberg estimated the core damage frequency of all operational reactors in the 1980s to be $15 \times 10^{-5}$ per reactor year, "within a factor three of the core melt probability" as estimated by the Rasmussen report (Spiewak and Weinberg 1985, p. 436).

In policy-making, an even higher probability of core melt down seems to have become acceptable in the years after, namely $10^{-4}$ per reactor year (NRC 1986). This probability corresponds to once in every ten thousand reactor years based, undoubtedly, on the number of reactors in operation in the 1980s (ca. 500) which thus meant that an accident would probably occur once in every 20 years.[1] However, serious growth was anticipated during what was known as the Second Nuclear Era in the 1980s; forecasts for as many as 5,000 reactors were made. Ten times more reactor years means that accidents could in principle happen ten times more frequently: a subsequent core melt accident probability of once every 2 years[2] was deemed unacceptable in terms of public confidence (Weinberg and Spiewak 1984). Safer nuclear reactors were therefore needed.

Most of the reactors in operation at the time of the TMI accidents were LWR-type reactors. They were originally designed for maritime purposes, the leading design criteria being compactness and simplicity rather than enhanced safety. As these reactors grew bigger for commercial energy production, the possibility of a core melt and its consequences became a serious challenge; i.e., new reactors had greater power capacities and more radioactive inventory. At that time several safety systems were then added to existing designs in order to enhance safety, but they also made the design "immensely complicated" because various additional "electromechanical devices, such as valves, scram rods, emergency pumps, and backup diesels" were then needed (Spiewak and Weinberg 1985, p. 432). This reliance on "external mechanical and/or electrical power, signals or forces" makes active intervention in the event of an incident necessary; the LWR designs of the early days therefore came to be known as "actively safe" (IAEA 1991, p. 10).

The first steps taken toward creating safer systems involved removing "[s]ome potential causes of failure of active systems, such as lack of human action or power failure" (IAEA 1991, p. 10); such systems came to be known as "passively" safe systems. The level of passivity of any system should be considered in terms of the number of external factors that have been removed. It would therefore be best to speak of higher or lower categories of passivity. IAEA illustrates this by giving an example. When a system's dependence on external power supply has been replaced

[1]Each year, 500 reactor years would pass, which means that based on the probability of $10^{-4}$, the expected number of accident would be $5 \times 10^{-2}$ (i.e., $500 \times 10^{-4}$) or simply once in every 20 years.

[2]Calculation: $5{,}000 \times 10^{-4} = 5 \times 10^{-1}$ or once in every 2 years.

by an internal power source, such as a battery, to supply active components, we can speak of the system being passive since it does not depend on a potential external failure, but this will probably be the lowest category of passivity (IAEA 1991); higher levels of passivity could be reached by removing reliance on more external factors, for instance, by removing reliance on power sources at all.

The first ideas on ways of making nuclear reactors safer focused on changes in the design of the LWRs; these changes were considered to be "incremental changes" that did not drastically affect the design philosophy (Firebaugh 1980). Below we will first discuss the rather small changes made in LWR design for the sake of safety. In contrast to these small changes, more substantial change is also conceivable. With these concepts, rather than adding incremental safety features, the design would emerge from a different safety philosophy, namely, that of *inherent safety*. Such drastic changes are hailed the transformation of the technological regime in the design of nuclear reactors (Van de Poel 1998).

The notion of inherent safety in reactor design shows some similarities with the design of chemical processes. Designing for inherent safety in chemical processes (and plants) has been introduced by Kletz (1978) and it entails that "we should avoid or remove hazards rather than add on protective equipment to control them" (Edwards and Lawrence 1993, p. 252). The same rationale has been adopted by the IAEA (1991, p. 9): that is, "inherent safety refers to the achievement of safety through the elimination or exclusion of inherent hazards through the fundamental conceptual design choices made for the nuclear plant." Before a power plant can be declared completely inherently safe, all these hazards have to be eliminated, but that is simply not feasible. We therefore speak, instead, of degrees of inherent safety. Thus, when in a reactor an inherent hazard is eliminated, it is inherently safe with respect to that specific hazard; for instance, when no combustible materials are used, a reactor is inherently safe against fire regardless of whatever incident or accident might occur (IAEA 1991). One important piece of rationale behind reactor design is the notion that new reactors should be made inherently safe in terms of being resistant to meltdown or core damage; more will be said about this below.

## Nuclear Fuel Cycles

Producing electricity requires more than a nuclear reactor to supply heat to a turbine. There are many steps required prior to electricity production (front end) and after reactor operation (back end). The whole process is called a *nuclear fuel cycle* and it starts with the mining and milling of uranium ore and ends with the possible treatment of the waste product and its geological disposal. There are now two ways to produce nuclear power, through open and through closed fuel cycles. Both methods use a LWR and both use uranium as fuel. Natural uranium contains two main isotopes, i.e., $^{235}U$ and $^{238}U$. Only the first isotope ($^{235}U$) is fissile and is used in LWRs as fuel, but it only constitutes 0.7 % of all natural uranium; this is why uranium is *enriched* by which we increase the fraction of the fissile isotope $^{235}U$ to 3–5 % for energy production in LWR. Irradiating uranium produces other

materials, including plutonium ($^{239}$Pu) and other fissile and non-fissile plutonium isotopes as well as *minor actinides*. Actinides are elements with similar chemical properties. Uranium and plutonium are the major constituents of spent fuel and so they are known as major actinides. Neptunium, americium, and curium are produced in much smaller quantities and are thus termed minor actinides. Fission products are a mixture of radionuclides that will decay to a nonhazardous level after approximately 250 years.

In the open fuel cycle, an isotope of uranium ($^{235}$U) is *fissioned* – split – in the reactor. The *spent nuclear fuel* is then designated for disposal underground and will take 200,000 years to become stable. The required storage time is dominated by plutonium. As stated above, less than 1 % of the uranium ore consists of the fissile isotope $^{235}$U. The major isotope of uranium ($^{238}$U) is non-fissile and needs to be converted into a fissile material for energy production: plutonium ($^{239}$Pu). In the closed fuel cycle, spent fuel undergoes a chemical process to separate useable elements, including the not irradiated uranium fuel as well as the plutonium produced during irradiation; this chemical treatment is referred to as *reprocessing*. During reprocessing, uranium and plutonium isotopes in the spent fuel are isolated and recovered. Recycled uranium could either be added to the beginning of the fuel cycle or used to produce *Mixed Oxide Fuel* (*MOX*) that is used as fuel in some nuclear reactors. The waste remaining after reprocessing is referred to as high-level waste (HLW), and it has a radiotoxicity higher than that of natural uranium for approximately 10,000 years dominated by the minor actinides.

## Values in Nuclear Engineering Design

Values are relevant to many of the choices that we make, also with regard to the design of technology; they reflect our understanding of the rightness and wrongness of those choices. The term value indeed has definitions that extend beyond philosophy and ethics. That said, the focus of this chapter is confined to the moral values that deal with how we want the world to be. We should not, however, confuse values with the personal interests of individuals; values are the general convictions and beliefs that people should hold paramount if society is to be good (Van de Poel and Royakkers 2011). Indeed, "the good" might be perceived differently by different individuals. In the following paragraphs, we will give definitions of these values, as they have been presented by the relevant international nuclear organizations. We believe that contention often arises more from how different values should be ranked in terms of their importance (moral or otherwise) than from how a single value is conceived of.

### Safety

As mentioned earlier, safety has played a key role in the developments of civilian nuclear technology; the detrimental health impacts of ionizing radiation were known long before the deployment of nuclear power in the 1950s (Clarke and Valentin 2009). The notion of safety is sometimes used in absolute terms (safety as

an absolute, as equated to no harm) and sometimes in relative terms (safety in terms of reducing the possibility of harm). Due to the many uncertainties we deal with in engineering design, safety is often interpreted in relative terms (Hansson 2009). This is certainly the case when addressing radiation risk, particularly since it is the accumulation of ionizing radiation that can have health impacts (see also the chapter "▶ Design for the Value of Safety" in this volume). The philosophy of radiation protection is "to reduce exposure to all types of ionizing radiations to the lowest possible level" (ICRP 1959, p. 10). The underlying rationale is that we reduce the level of radiation such that we eliminate or at least reduce the probability of detrimental effects. So, the "health objective" prescribes that the "deterministic effects are prevented, and the risks of stochastic effects are reduced to the extent reasonably achievable" (Valentin 2013, p. 19).

In short, safety as a value refers here to those concerns which pertain to the exposure of the human body to radiation and its subsequent health effects.

Due to the longevity of nuclear waste, safety is a value that relates to future generations as well. The safety of future generations has been one of the concerns from the early days of nuclear power production. The Nuclear Energy Agency states that we should offer "the same degree of protection" for people living now and in the future (NEA-OECD 1984). The IAEA reiterates this in its Safety Principles where it states that nuclear waste should be managed in such a way that "predicted impacts on the health of future generations will not be greater than relevant levels of impact that are acceptable today" (NEA-OECD 1995, p. 6).

### Security

In the IAEA's Safety Glossary, nuclear security is defined as "any deliberate act directed against a nuclear facility or nuclear material in use, storage or transport which could endanger the health and safety of the public or the environment" (IAEA 2007, p. 133). One can argue that "security" as defined here also refers to the safety considerations discussed above. We shall, however, keep the value of "security" separate in our analysis so as to be able to distinguish between unintentional and intentional harm. We define "security" as the protecting of people from the malicious intentional harmful effects of ionizing radiation resulting from sabotage or proliferation. Thus security variously relates to nuclear theft and unauthorized access, to the illegal transfer of nuclear material or other radioactive substances at facilities (IAEA 2007, p. 133), and also to the dissemination of technical know-how or facilities that could lead to the proliferation of nuclear weapons. Proliferation threats arise either from using highly enriched uranium (HEU) which has been enriched up to 70 % (and higher) or from producing or separating weapon-grade plutonium in reprocessing plants; more will be said about this in section "Design for Nuclear Values."

### Sustainability

Sustainability is one of the most discussed and perhaps most contested notions in the literature on nuclear power. It is not our intention to enter into those discussions here and certainly not to assess the degree of sustainability of nuclear power.

One common and influential definition concerning sustainable development is the Brundtland definition that emphasizes the ability of present generations to meet their own need without compromising the ability of future generations to meet their needs (WCED 1987; see also the chapter "► Design for the Value of Sustainability" in this volume). In nuclear power production and nuclear waste management, this definition at least relates to two specific issues, namely, the state of the environment as posterity bequeaths from us – referred to as *environmental friendliness* – and the availability of natural (nonrenewable) energy resources on which future generations' well-being relies, referred to as *resource durability*.

### Environmental Friendliness

The value of *environmental friendliness* relates to the accompanying radiological risks to the environment. Radiological risks, as perceived in this chapter, express the possibility or rather probability that radioactive nuclides might leak into the biosphere and harm both people and the environment. Issues that relate to the harming of human beings have already been subsumed under the heading safety. The effect of the same radiation on the environment and nonhuman animals is subsumed here under the heading of environmental friendliness.

Whether we should protect the environment for its own sake or for what it means to human beings is a long-standing and still ongoing discussion in the field of environmental philosophy. In the anthropocentric (human-centered) approach, this notion would solely encompass those aspects of the environment that are relevant to human health. The non-anthropocentric approach would address the consequences of radiation in the environment without making reference to what this means for human beings.

Various UN policy documents, including IAEA publications, interchangeably refer to both approaches. We do not intend to take a stance on this matter either. We preserve the value of "environmental friendliness" as a separate value in order to allow for a broader number of views to be reflected with this set of values. Those who would follow the anthropocentric approach will then simply merge this value with the value of "safety."

### Resource Durability

If we now consider the period from the industrial revolution up until the present, it would be fairly straightforward to conclude that the availability of energy resources has played a key role in achieving (and sustaining) people's well-being. The appropriate consumption of nonrenewable natural resources over time is one of the central issues of sustainability; "later generations should be left no worse off [. . .] than they would have been without depletion" (Barry 1989, p. 519). Since it would be irrational to expect the present generation to leave all nonrenewable resources to its successors and since replicating such resources is not an option either, it has been argued that we need to offer compensation or recompense for depleted resources "in the sense that later generations should be no worse off [. . .] than they would have been without depletion" (Barry 1989, p. 519). The value of *resource durability* is therefore defined as the availability of natural resources for the future or as the providing of an equivalent alternative (i.e., compensation) for the same function.

### Economic Viability

The next value that we shall discuss in relation to sustainability is that of *economic viability*. One might wonder whether economic issues have inherent moral relevance and whether it is justified to present economic durability as a moral value. We can safely assume that the safeguarding of the general well-being of society (also, for instance, including issues of health care) has undeniable moral relevance. However, in our understanding of economic viability in this chapter we do not refer to general well-being but only to those aspects of well-being that have to do with nuclear energy production and consumption. With this approach economic aspects do not have any inherent moral relevance; it is what can be achieved through this economic potential that makes it morally worthy.

This is why we present the value of economic viability in conjunction with other values. First and foremost, economic viability should be considered in conjunction with resource durability. In that way it relates to the economic potential for the initiation and continuation of an activity that produces nuclear energy. As we shall see in the following sections, some future nuclear energy production methods might require serious R&D investments for further development; particularly those new methods that are based on new types of reactors which would require serious investment prior to industrialization. Economic viability could also become a relevant notion when we aim to safeguard posterity's safety and security by introducing new technology for the reducing of nuclear waste lifetime. In general, economic viability is defined here as the economic potential to embark on a new technology and to safeguard its continuation in order to maintain the other values.

### Intergenerational Justice

Concerns about depleting the Earth's resources and damaging the environment have triggered a new debate on the equitable sharing of goods over the course of generations; this is referred to as intergenerational justice. The main rationale is that we should consider justice in what we leave behind for generations to come after us. There are two ways in which intergenerational justice relates to nuclear power production and to waste management. First of all, assuming that this generation and those that immediately follow will continue depleting uranium, a nonrenewable resource, there will be evident intergenerational justice considerations to bear in mind. Secondly, the production of nuclear waste, and its longevity in terms of radioactivity, signifies substantial present benefits with deferred costs. In nuclear waste management this notion of justice across generations has been influential, particularly in promoting geological repositories as final disposal places for nuclear waste.[3] Also in designing nuclear reactors and their surrounding systems, intergenerational justice has proven to be a relevant value. This will be elaborated in section "Design for Nuclear Values."

[3]This subsection is mainly drawn from the following publication, in which the role of intergenerational justice in nuclear waste management has been extensively discussed (Taebi 2012).

## Design for Nuclear Values

Section "Nuclear Technology" briefly presented the development of nuclear reactors design and introduced four main values that play a key role in designing reactors and their surrounding systems, such as nuclear fuel cycles. In this section, we will first operationalize these values by specifying how they relate to different phases of the nuclear fuel cycle. In this way we can assess fuel cycles based on the presented values. More importantly, we will focus on how these values have played a role in the design of nuclear technology, both in designing new fuel cycles and the associated nuclear reactors.

## Designing Nuclear Fuel Cycles

In the interests of brevity, we will not elaborately discuss the operationalization of these values for the assessment of the two fuel cycles, but we shall briefly explain how this could be effected.[4] What is particularly important in this analysis is that we link the impact of different steps in the fuel cycle to the values presented and evaluate the extent of those impacts. In so doing we should distinguish between the impacts for both the present and future generations. Let us illustrate this by taking an example in which we shall operationalize the value "safety."

When assessing safety issues in an open fuel cycle, we should at least address the following steps that relate in one way or another to the safety issues: (1) mining, milling, enrichment, and fuel fabrication; (2) transport of (unused) fuel and spent fuel; (3) reactor operation and decommissioning period; (4) interim storage of spent fuel; and (5) final disposal of spent fuel in geological repositories. The open fuel cycle is represented by the thick (black) lines in Fig. 1.

Each of the five aforementioned steps relates to a different time period. This means that they would affect the interest of the present and future generations differently. Most steps would last for the period of reactor operation or maximally for several decades after that particular period, for instance, for the decommissioning of the reactor and for the interim storage of the waste. The final disposal of waste obviously has an impact for a much longer period of time. From the perspective of long-term safety concerns, there will be potential burdens after spent fuel has been placed in the geological repositories; these concerns will potentially last for the life-time of the spent fuel or approximately 200,000 years. So this is the period in which the value of safety is potentially at stake. Figure 2 shows the result of such analysis for the open fuel cycle.

In this way we can evaluate the existing fuel cycles based on the values and how they have been operationalized to relate to specific steps in the fuel cycle. Elsewhere we argued that each fuel cycle would promote certain values and sometimes,

[4]For an elaborated discussion of the operationalization of the values in fuel cycles, see (Taebi and Kadak 2010).

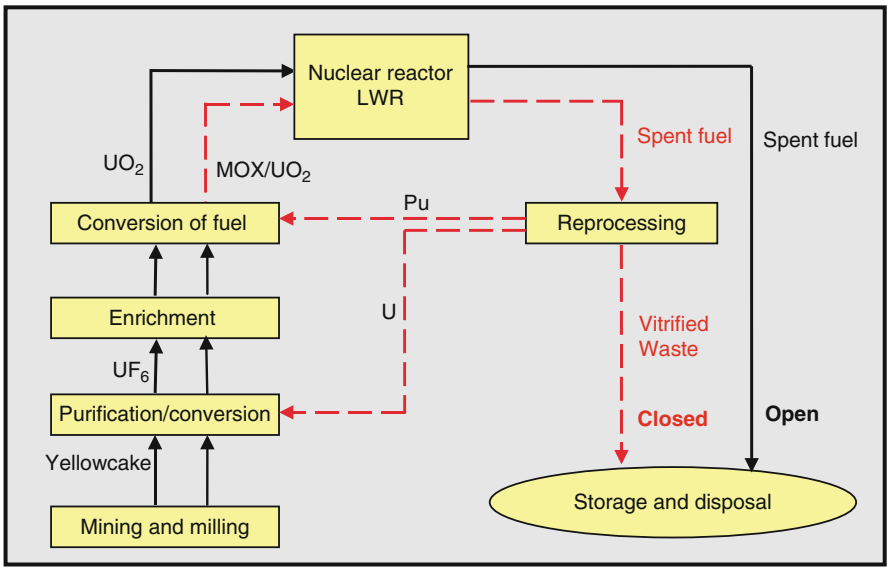



**Fig. 1** An overview of the open and closed nuclear fuel cycle; the thicker (*black*) *lines* and *arrows* represent the open fuel cycle, while the thinner (*red*) ones the closed fuel cycle (Source: Fig. 2 in Taebi and Kloosterman (2008))

as a result, undermine other values. So choices need to be made between these values. So, the open fuel cycle seems to be preferable from the perspective of present generations, since it creates fewer safety and security risks and is less costly. The closed fuel cycle is, on the other hand, more beneficial from the point of view of future generations, because it reduces the long-term safety concerns of waste disposal while helping to extend nonrenewable resources farther into the future. At the same time, the closed cycle creates more short-term safety and security concerns and economic burdens. The choice of a given fuel cycle should thus be made by weighting the moral relevance of each values in a temporal sense (Taebi and Kloosterman 2008).

This ex-post analysis shows that when policy-makers opt for a specific fuel cycle, these value trade-offs are made implicitly. This analysis gains more relevance when we include the values in ex ante analysis of what we deem to be a desirable future nuclear technology. This approach accounts for values through the design process and it is referred to as Value Sensitive Design (Friedman 1996). It constitutes an attempt to uphold human values with ethical importance as design criteria, so that we can proactively think and guide future technologies (Friedman and Kahn 2003).

Let us now take the following example to illustrate this point. As mentioned above, the waste emanating from the open and closed fuel cycles is radiotoxic for either 200,000 or 10,000 years. Societies might find it desirable to further reduce the waste lifetime in order to enhance the value of long-term safety. If we were to

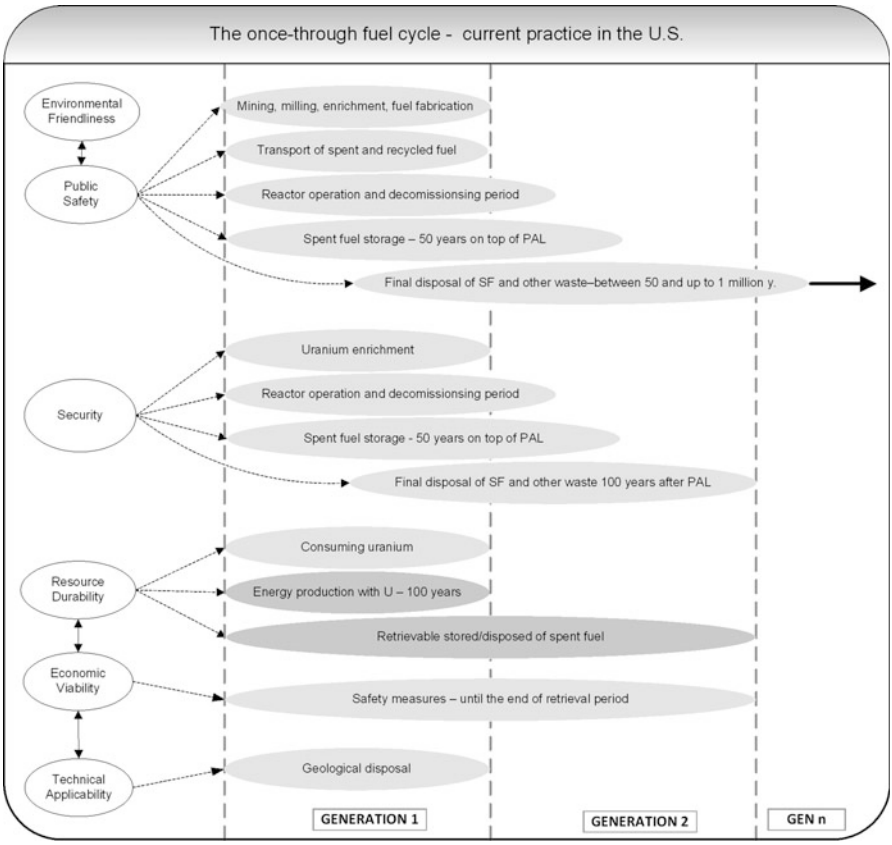



**Fig. 2** Relating values to concrete consequences and to the associated *Period in which the Activity Lasts* (PAL) as seen in a once-through fuel cycle or the current practice in the United States. The *light* and *dark gray* ellipses represent the respective burdens and benefits. The horizontal *black arrow* depicts a projection of certain considerations extending into the future and far beyond the time frame of the charts (Source: Fig. 3 in Taebi and Kadak (2010)). (In this paper 100 years is taken as the definition of the present generations. "[T]he immediately following generation as everyone who is now alive, including the infants born in the last couple of moments, then it will be a much longer period of time – namely, the length of people's average life expectation – before the current generation ceases to exist and we can speak of a future generation" (Taebi and Kadak 2010, p. 1345); building on De-Shalit (1995))

incorporate such societal desire into technological development, there is one waste management method that would be particularly interesting, namely, the partitioning and transmutation (P&T) method. It could in principle reduce the waste lifetime by approximately a factor of 20 to 500–1,000 years. Its feasibility has already been shown at lab level, but the relevant technologies surrounding multiple reprocessing and fast reactors still need to be further improved. Fast reactors – accelerator-driven systems (ADS) could alternatively be used – are applied to create higher energy neutrons, which are capable of fissioning a greater number of isotopes, including

the minor actinides, in the reprocessed spent fuel. This would help reduce the waste lifetime (IAEA 2004). Proactive thinking in terms of the values at stake could then help determine how to incorporate the value of safety, long term and otherwise, in nuclear waste management while elaborating on the implications for the other values at stake. P&T creates additional burdens for present generations in the form of the safety issues derived from additional nuclear activities, security issues emerging from the multiple recycling of plutonium, and economic burdens for the further development of the technology, including the required R&D funding. This brings us back to the fundamental question of which value should be preferred and for what reasons. More specifically, can additional burdens upon present generations sufficiently be justified?[5]

In sum, in this subsection we argued intergenerational justice is inevitably an important value when we are to choose between different fuel cycles. We further showed that in answering the thorny question of "what justice exactly entails for future generations," we need to assess the impacts of fuel cycles in terms of the aforementioned value, namely, safety, security, sustainability, and economic viability.

## Designing Nuclear Reactors

In the remainder of this section we will focus on the role values have played in nuclear reactor design. Our focus is on new generations of nuclear reactors. We will take the already operational Gen II reactors as the default situation and show how the value of safety has influenced the design of Gen III and Gen III reactors. Design changes could either be through incremental changes (compared to LWRs) or more radical changes. Table 1 presents a list of the reactors discussed in this section and summarizes the type of proposed changes and indicates the assigned probabilities of core damage. We will also focus on how other values such as sustainability are becoming increasingly important in the design of nuclear reactors and how that changes proposed designs.

### Gen III: ABWR

Gen III is the evolutionary successor to LWRs bringing design improvements in fuel technology, thermal efficiency, and, most importantly, passive safety systems; advanced designs in both BWR and PWR are introduced in this generation (Goldberg and Rosner 2011, p. 6). Only four Gen III reactors are currently operable, all advanced boiling water reactors (ABWRs). The safety improvements in this type of reactor, compared to the BWR, include the addition of ten separate internal pumps at the bottom of the reactor vessel, the addition of several emergency cooling systems, and the encasing of the reactor vessel in thick fiber-reinforced concrete containment. These incremental changes helped to simplify the design while

[5]Please see for an elaborate discussion of this issue (Taebi 2011).

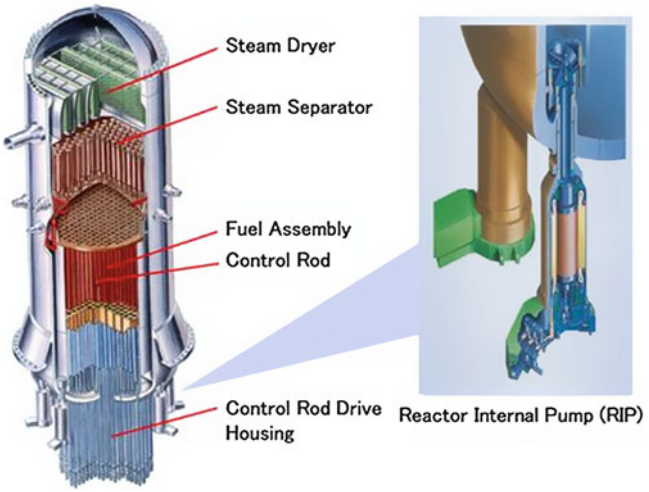



**Fig. 3** The reactor pressure vessel of ABWR and the magnified internal pumping system (Source: http://nuclearstreet.com/nuclear-power-plants/w/nuclear_power_plants/abwr-ge-hitachi.aspx)

simultaneously improving performance. For instance, the situating of water pumps a short distance away from the reactor vessel would eliminate the need for complex piping structures (as with the BWR) and thus simplify the cooling system, while the presence of multiple pumps would increase safety in the event of failure of one or more of the pumps (see Fig. 3). The ABWR is designed and manufactured by the General Electric Hitachi Nuclear Energy (GEH) company in conjunction with Toshiba. The manufacturers anticipate a core damage frequency of $1.6 \times 10^{-7}$, or approximately once in five million reactor years, which is 300 times less probable than the original BWR, as calculated in the Rasmussen Report (namely, $4.6 \times 10^{-5}$); see also noted in Table 1.

## Gen III+: AP1000 and HTR-PM

The most significant improvement in Generation III+ reactors is the inclusion of "some designs of passive safety features that do not require active controls or operator intervention but instead rely on gravity or natural convection to mitigate the impact of abnormal events" (Goldberg and Rosner 2011, p. 8). The improvements in Generation III could either qualify as incremental changes or as radical changes to the existing designs. We will discuss examples from both categories, i.e., AP1000 (a successor to the PWR) and a high-temperature reactor pebble-bed module (HTR-PM) which may be seen as a radically new reactor that takes safety as its primary design criterion and starts from scratch.

The core, reactor vessel, and internals of AP1000 are based on the conventional PWRs, built by Westinghouse. AP1000 is clearly a more passive safety system by using fewer safety valves and pumps and less safety piping and cables; furthermore it employs a (passive) core cooling system with three sources of water "to maintain

cooling through safety injections," passive residual heat removal, and (passive) containment cooling system to "provide the safety-related ultimate heat sink for the plant" (Schulz 2006, pp. 1551–1552). According to the manufacturer, these passive systems would reduce the probability of a core melt in an AP1000 to $4.2 \times 10^{-7}$, making it almost 200 times less probable than the *original* PWR (i.e., according to the Rasmussen Report $2.6 \times 10^{-5}$).

A more radical change in reactor design came with the introduction of HTR-PM. This type of reactor was first built in Germany – AVR (Arbeitsgemeinschaft Versuchsreaktor) – and further developed in South Africa, PBMR (Pebble-Bed Modular Reactor). Developments are now being continued in China. HTR-PM takes various safety and economic goals as its primary design criteria; for instance, the design of the reactor should not require "anyone living near the site boundary to take shelter or be evacuated" following any internal event in the reactor or an external event affecting the condition of the reactor (Koster et al. 2003, p. 232). This safety criterion has further been translated into an economic goal from the point of view that these reactors do not need large exclusion zones for operational purposes. That is again beneficial when it comes to licensing matters and to transporting electricity to populated areas. This safety regime has been termed inherently safe by the IAEA (1991, p. 9). The HTR-PM is further designed to offer levels of radiation safety to workers that would be higher than the international recommended standards (Koster et al. 2003).

To accomplish such levels of safety, two important design changes have been proposed in order to ensure that the reactor does not overheat and to make the fuel resistant to heat. The first change concerns the shape of the reactor. It will be a long cylinder with a small radius. This facilitates natural heat exchange with the environment due to the large reactor surface area. The heat is then transported to the cooling system which has the capacity to passively absorb this heat for more than 72 h (Koster et al. 2003, p. 236). It is important to observe that this safety improvement has an adverse effect on the economic aspects of the reactor because during normal operation, a part of the neutrons sustaining the fission chain reaction will leak out of the core, requiring a slightly higher enrichment of the uranium fuel.

The second change in the design concerns the revolutionary approach to fuel and its cladding; see Fig. 2. HTR-PM fuel consists of fuel spheres containing small coated particles. Each particle consists of a small amount of uranium oxide (i.e., fuel) which is encompassed in four layers of coating. Especially having two layers of pyrolytic graphite and one layer of silicon carbide (SiC) would make leakage of radioactive nuclides (i.e., fission products) substantially less probable, since those layers can withstand very high temperatures and can thus support the integrity of fuel spheres.[6]

[6]This paragraph is partly based on information provided by the South African company, Pebble-Bed Modular Reactor (Pty), that built PBMR. See: http://www.pbmr.com/contenthtml/files/File/WhynoChernobyl.pdf.

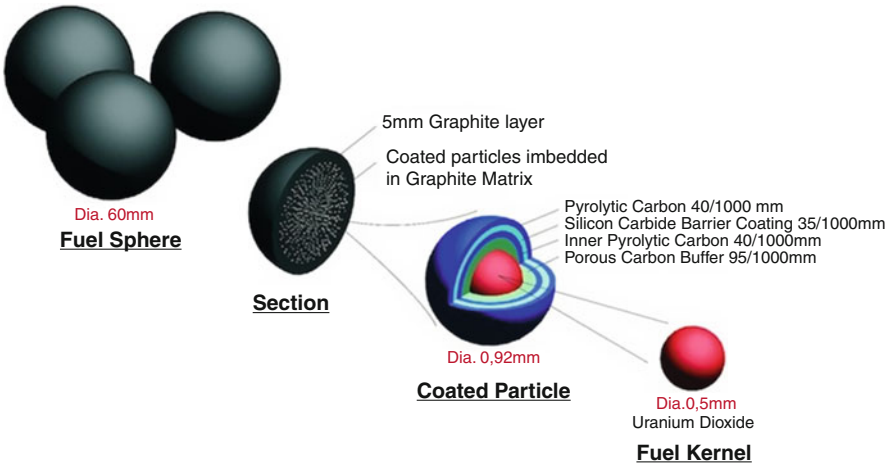



**Fig. 4** The fuels of an HTR-PM reactors

In conjunction with these design characteristics, a core meltdown would – in principle – be ruled out in an HTR-PM. In the probabilistic risk assessments for this type of reactor, one looks instead at the possibility of radionuclides being released into the environment in the event of damage occurring to the SiC coating, after which radionuclides could migrate from the fuel particles through the graphite to the coolant (Koster et al. 2003, p. 232). This could occur at a temperature above 1,600 °C or after chemical degradation of the fuel resulting from a large ingress of water or air (the coolant is inert helium) in the fuel at a temperature of 1,200 °C. The probability of radionuclide release in a modular high-temperature gas-cooled reactor is $5 \times 10^{-7}$; the released doses in such cases are, however, expected to be so low that sheltering would not be required (Silady et al. 1991, p. 421; Fig. 4).

## Gen IV Fast Reactor: GFR and MSR

The latest developments in reactor technology are concentrated in Gen IV reactors which are designed to reconcile many different criteria. Firstly, these reactors should help us deploy the major isotope of uranium, the non-fissile $^{238}$U, thus enhancing resource durability in order to meet the value of sustainability. One must bear in mind that less than 1 % of all naturally occurring uranium is deployable in conventional thermal reactors, while fast reactors are capable of converting the major isotope of uranium ($^{238}$U > 99 %) to fissile $^{239}$Pu. These reactors are the *breeder* reactors that breed (or make) new fuel (i.e., $^{239}$Pu). During operation this plutonium isotope can be used again as fuel. Other types of breeder reactors could be designed to use the more naturally abundant thorium as a fuel. This kind of reactor, the molten salt-cooled reactor (MSR), will be discussed here. Apart from meeting sustainability requirements, Generation IV reactors are intended to furthermore enhance long-term safety by reducing the volume and lifetime of nuclear waste. The gas-cooled fast reactor (GFR) will now be discussed.

The gas-cooled fast reactor is a fast-spectrum helium-cooled reactor that is designed to make efficient use of the major uranium isotope, but it is also designed with a view to waste management. High-energy fast neutrons enable this reactor to irradiate isotopes that thermal neutrons in conventional thermal reactors (e.g., LWR) cannot fission. This has evidently sustainability benefits for the durability of uranium as an energy source; the term "plutonium economy" refers to the implementation of fast reactors for energy generation purposes.

The second rationale behind introducing GFR is to eliminate the troublesome actinides which, again, thermal neutrons cannot fission. Partitioning and transmutation (P&T) as discussed in section "Design for Nuclear Values" requires the use of a fast reactor. P&T deals essentially with spent fuel recycling which is, in principle, the same technology as that currently used in closed fuel cycles. However, this type of reprocessing of fast reactor spent fuel is a technology that needs to be further developed for the recycling of actinides (Abram and Ion 2008). In addition, the expected result in terms of waste lifetime reduction can only be achieved after multiple recycling, and so therefore, it is recommendable to build such a reprocessing plant onsite in order to avoid the extensive transporting of spent and *fresh* fuel. A reprocessing plant is, however, only economically viable if it is built for many reactors. This means that a P&T cycle based on multiple recycling is only thinkable if several fast reactors (e.g., GFRs) and a fast reactor fuel reprocessing plant are present on the same site. Such a set up would introduce further proliferation concerns because of the continuous flow of plutonium in the cycle. The GFR requires further R&D development in the areas of fuel, the fuel cycle process, and its safety systems. The high core power density requires additional safety devices and systems, but the design must guarantee that the need for active safety systems is minimized (DOE 2002, p. 25; Fig. 5).

Molten salt-cooled reactors are probably the most ambitious kind of Gen IV reactors since they seriously depart from conventional reactor technology. The MSR was first proposed as a US aircraft propellant and it is one of the few reactors that can use naturally abundant thorium as a fuel. This reactor will run on a combination of uranium-thorium fluoride mixed in a carrier salt such as beryllium fluoride and lithium fluoride. This salt – that serves both as the fuel and the coolant – will continuously circulate through the reactor core and the heat exchanger to transfer the heat to a second circulation system for electricity production. A part of this fuel/coolant will then go to a chemical processing plant where the fission products will be removed and new fissile material will be added. "This continual processing of the fuel allows operation without refueling outages, and the fluid fuel offers a unique safety feature where the entire fuel inventory can be drained from the reactor in the event of an accident" (Abram and Ion 2008, p. 4328); see the emergency dump tanks in Fig. 3. The combination of corrosive and highly radioactive salt constantly running through the reactor places serious and extreme requirements on the material performance in the piping of the primary circuit and the equipment of the processing plant. Among other technical challenges, serious R&D effort needs to be put into fuel development, molten salt chemistry control, and corrosion study carried out on the relevant materials (DOE 2002, pp. 34–35).

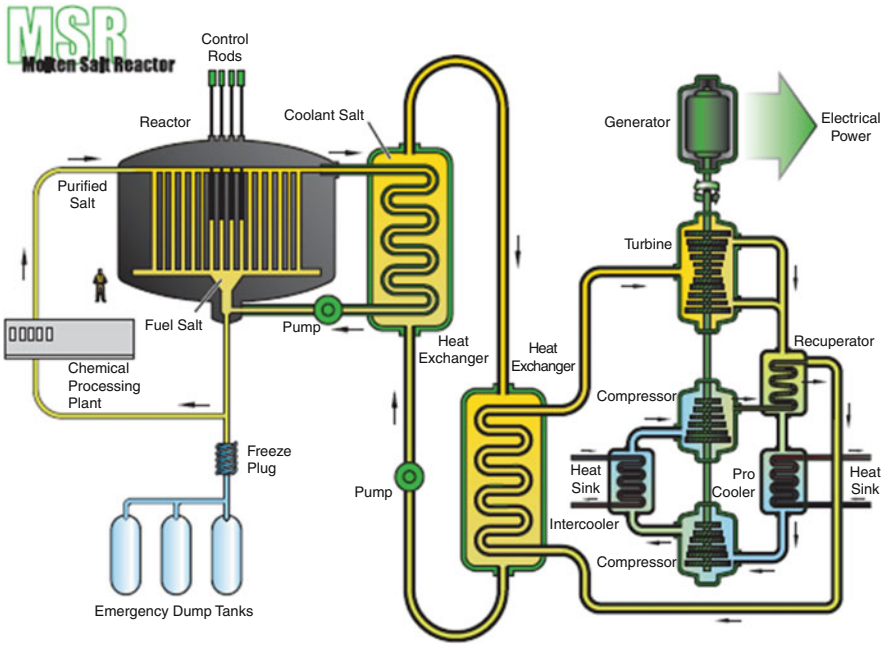



**Fig. 5** Molten salt reactor with a primary and a secondary circuit (Source: (DOE 2002, p 33))

Some people maintain that the MSR lies at the boundary of Gen IV technology and is perhaps too ambitious to be industrialized (Abram and Ion 2008).

## Open Issues and Future Work

Future endeavors should at least focus on three open issues that will be discussed in this section.

### The Inability of the Probabilistic Risk Assessment

The safety of nuclear reactors has been systematically conceptualized since the introduction of probabilistic risk assessment (PRA) as proposed by the Rasmussen group in 1975. This was based on a fault-tree analysis that examined the undesired events and assigned probabilities to each event. Since the Rasmussen Report, the probability of core damage has been the leading criterion in safety studies. Even though this method has clear advantages such as highlighting weaknesses, different probabilities assigned in the literature are not necessarily referring to the same types of events; we cannot therefore always easily compare risks in terms of the calculated probabilities. For example, core damage is a different concept than core

meltdown; so we cannot easily compare those two risks in terms of their probabilities. More importantly, without a meltdown radioactive nuclides can also leak into the environment following core damage while having a meltdown does not necessarily mean that there will be leakage into the environment, if the containment of the reactor retains its integrity. When we consider the uncertainties regarding the health impacts of the different types of nuclides and radiation, the complexity of the matter reveals itself in all its glory. To sum up, probabilistic risk assessment is a very good indication of safety, but it is not the final word in the discussion on reactor safety. Yet, PRA is absolutely indispensable when assessing the safety of nuclear reactors and comparing different reactors, even though the accuracy of the current estimation could be questioned in the light of the recent events in Fukushima (Goldberg and Rosner 2011); see also (Taebi et al. 2012, pp. 202–203).

Safer reactors could be realized by including incremental changes in the conventional current light water reactors. These reactors, of which the pressurized water reactors were originally designed to be used in the maritime sector, had simplicity and compactness as their main design criteria. When they were deployed on a large scale for commercial energy production purposes, many safety systems such as valves, emergency pumps, and backup diesels were added to the original design. The paradox of these safety systems was that they simultaneously made the reactors immensely complex and, in the process, unsafe. Incremental changes were proposed to remove these complexities. Reactors could furthermore be made passively safe by means of incremental changes; passively safe reactors reduce the need for human intervention and other external systems, for instance, through the use of emergency cooling systems that are solely based on gravitation, all of which reduce the probability of core damage. Safety could be further improved by bringing radical change to reactor design, for instance, by introducing inherently safe reactors that eliminate inherent hazards through fundamental design changes.

## Fukushima and the Future of Safety

As stated earlier, safety has always been one of the important driving forces behind serious changes in reactor design philosophy. It is particularly major nuclear accidents that seem to have affected people's thinking about reactor safety. It was, for instance, the Three Mile Island accident that initiated thinking about passively safe reactors and reducing the influence of operator action. It is now interesting to anticipate how the Fukushima Daiichi accidents might affect the design of nuclear reactors. We maintain that the proposed changes for the next couple of years will probably be incremental in two different respects. Firstly, the protection of the surrounding reactor systems, that is to say, the primary system of all but the oldest reactors in Fukushima withstood both the earthquake and the tsunami in 2011. The damage was caused by the defective external cooling system that was not as well protected as the reactor, all of which accelerated the accidents.

Secondly, changes can be expected in the cladding of nuclear fuel. Current nuclear fuel cladding is composed of different zirconium alloys, because of their favorable mechanical and physical properties. An important drawback of metallic cladding is, however, that it can undergo a reaction with water above a certain temperature. This chemical reaction generates hydrogen gas which could, in turn, cause a hydrogen explosion like that seen in the Fukushima reactors. A move toward ceramic cladding is thus to be expected. Reactors are further expected to be made less vulnerable to large external events.

## Designing for Conflicting Values

With the introduction of Gen IV reactors, sustainability also became a particularly relevant criterion in design. Indeed, security and economic viability have always played a role in design. We can assert that the Best Achievable Nuclear Reactor would maximally satisfy all these criteria, but as we have seen in the preceding sections, the safest reactor is not always the most sustainable one, while the reactor that best guarantees resource durability could easily compromise safety and security. Since we cannot meet all these criteria simultaneously, choices and trade-offs need to be made.

We highlighted these choices by discussing three promising future reactor types. So, depending on which design criteria will be decisive, drastically different reactors could be proposed. The high-temperature reactor pebble-bed module (HTR-PM) scores best on the criterion of safety because of the radical change it makes to safety design philosophy; core melt down is physically impossible in such a reactor; only core damage can occur. On the other hand, the molten salt-cooled reactor (MSR) scores best on resource durability because it can use naturally more abundant thorium as its fuel; however, this reactor type, until proven differently, scores low on chemical safety, because the highly radioactive liquid fuel is also chemically corrosive. The gas-cooled fast reactor (GFR) also scores high on sustainability since it uses the major isotope of natural uranium, but a GFR would score lowest on security since there is constantly plutonium in its cycle.[7] The relatively low score of MSR on security is attributable to the production of a certain isotope of uranium ($^{233}$U) that could be used for weapon purposes (no enrichment required).

In conclusion, when we aim to design for one single value, other values (or design criteria) will change simultaneously. This raises the question of to

[7]There are two remarks that need to be made. Firstly, it is the authors' opinion that an MSR would score the best on the sustainability criterion. This is because of the natural abundance and good dispersal of thorium compared to uranium. Secondly, the economic viability is based on a rough estimation made by the authors in which assumptions have been made with regard to the required research funding for the industrialization of these three reactors. HTR-PM with a prototype reactor in China seems to be the farthest ahead in its research, which makes it score best on economic viability, while MSR presumably still requires substantial research.

**Table 2** The conflicting design criteria for three important designs for the future of nuclear reactors

| | HTR-PM | GFR | MSR |
|---|---|---|---|
| **Safety** | ++ | − | + |
| **Security** | + | − − | − |
| **Sustainability (durability)** | − | + | + + |
| **Economic viability** | + | 0 | − |

This is merely an internal comparison based on each criterion or value (Assigning plusses and minuses as a means of internal comparison based on value does not imply that we can quantitatively compare these values. In other words, we cannot sum up the plusses and minuses for each reactor to see which one scores best. The only conclusion we can draw would be based on single criteria)

what extent we can compromise one design criterion for the achievement of another? We can rephrase this question in more general terms: to what extent could we jeopardize one value for the achievement of another value? Table 2 shows these conflicts in an internal comparison between the three reactors discussed here and on the basis of each criterion.

## Conclusions

In this chapter we introduced five main values that play a crucial role in nuclear design namely, safety, security, sustainability (both in terms of environmental friendliness and resource durability), economic viability, and intergenerational justice. We first elaborated on how each of these values has been perceived in the six-decade-old history of nuclear power and what role they have played in designing nuclear reactors and nuclear fuel cycles.

The main focus of the chapter was on incorporating these values into an ex ante analysis of what we deem to be a desirable future technology. This represents an approach that accounts for human values through the design process and it is referred to as Value Sensitive Design. Thinking in terms of values has already motivated the development of new fuel cycles such as partitioning and transmutation (P&T) and the design of reactors. These developments have however often been focused on one single value; also in nuclear technology safety has been a leading value. VSD aims to proactively balance different values in the design process. In balancing these values, choices need to be made between the values that we find to be important in design. These choices often go back to a fundamental issue that should be addressed proactively and prior to further development of the reactors. Research and R&D funds are scarce and previous experience shows that once policy-makers invest in a certain option, they are not easily inclined to shift focus later on because of the initial investments. Therefore it is crucial to address these ethical value conflicts prior to the further development of each reactor.

**Acknowledgment** The authors wish to thank Ibo van de Poel as well as Daniela Hanea for their valuable comments.

## References

Abram T, Ion S (2008) Generation-IV nuclear power: a review of the state of the science. Energy Policy 36(12):4323–4330

Barry B (ed) (1989) The ethics of resource depletion. In: Barry B (ed) Democracy, power and justice, essays in political theory. Clarendon Press, Oxford, pp 511–525

BBC (2011) China and Bill Gates discuss nuclear reactor plan 2011. Cited 8 Dec 2011. Available from http://www.bbc.co.uk/news/technology-16085385

Clarke RH, Valentin J (2009) The history of ICRP and the evolution of its policies. In: ICRP (ed) Application of the commission's recommendations for the protection of people in emergency exposure situations. Annals ICRP 37(5). Elsevier, Oxford, pp 75–110

De-Shalit A (1995) Why posterity matters: environmental policies and future generations. Routledge, London/New York

DOE (2002) A technology roadmap for generation IV nuclear energy systems. GIF-002-00. U.S. DOE Nuclear Energy Research Advisory Committee and the Generation IV International Forum, Washington, DC

EC-DGXII (1994) Externalities of fuel cycles. "ExternE" project', working documents 1–9, European Commission. Directorate Generale XII – Science, Research and Development/ Joint Research Centre

Edwards DW, Lawrence D (1993) Assessing the inherent safety of chemical process routes: is there a relation between plant costs and inherent safety? Process Saf Environ Prot 71(B4): 252–258

Firebaugh MW (1980) Acceptable nuclear futures: the second ERA [ORAU/IEA-80-1 1(P)]. Tennessee Institute for Energy Analysis, Oak Ridge Associated Universities, Oak Ridge

Friedman B (1996) Value-sensitive design. Interactions 3(6):16–23

Friedman B, Kahn PH (2003) Human values, ethics, and design. In: Jacko J, Sears A (eds) Handbook of human-computer interaction. Lawrence Erlbaum, Mahwah, pp 1177–1201

Goldberg SM, Rosner R (2011) Nuclear reactors: generation to generation. American Academy of Arts and Sciences, Cambridge, MA

Hansson SO (2009) Risk and safety in technology. In: Meijers A (ed) Philosophy of technology and engineering sciences. Elsevier, Amsterdam, pp 1069–1102

IAEA (1991) Safety related terms for advanced nuclear plants. IAEA, Vienna

IAEA (2004) Technical implications of partitioning and transmutation in radioactive waste management. IAEA, Vienna

IAEA (2007) IAEA safety glossary, terminology used in nuclear safety and radiation protection. IAEA, Vienna

ICRP (1959) Recommendations of the International Commission on Radiological Protection: revised December 1954. ICRP publication 1, vol 1, Annual ICRP. Pergamon Press, Oxford

Keller W, Modarres M (2005) A historical overview of probabilistic risk assessment development and its use in the nuclear power industry: a tribute to the late Professor Norman Carl Rasmussen. Reliab Eng Syst Saf 89(3):271–285

Kletz TA (1978) What you don't have, can't leak. Chem Ind 6:287–292

Koster A, Matzner HD, Nicholsi DR (2003) PBMR design for the future. Nucl Eng Des 222(2):231–245

Leurs BA, Wit RCN (2003) Environmentally harmful support measures in EU member states. CE Delft, Delft. Report for DG Environment of the European Commission

Lilienthal DE (1980) Atomic energy: a new start. Harper and Row, New York

Minarick JW, Kukielka CA (1982) Precursors to potential severe core damage accidents: 1969–1979. A status report. US Nuclear Regulatory Commission (NRC), Oak Ridge National Laboratory

NEA-OECD (1984) Long-term radiation protection objectives for radioactive waste disposal, report of a group of experts jointly sponsored by the Radioactive Waste Management

Committee and the Committee on Radiation Protection and Public Health. Nuclear Energy Agency, Organisation for Economic Co-operation and Development, Paris
NEA-OECD (1995) The environmental and ethical basis of geological disposal of long-lived radioactive wastes: a collective opinion of the Radioactive Waste Management Committee of the Nuclear Energy Agency. Nuclear Energy Agency, Organisation for Economic Co-operation and Development, Paris
NRC (1975) In: Rasmussen NC (ed) Reactor safety study. An assessment of accident risks in U.S. commercial nuclear power plants. WASH-1400-MR; NUREG-75/014-MR. Nuclear Regulatory Commission, Washington DC
NRC (1986) Safety goals for the operations of nuclear power plants: policy statement, republication. 51 FR 30028. Nuclear Regulatory Commission (NRC), Washington, DC
NRC (1994) Final safety evaluation report related to the certification of the advance boiling water reactor design, main report, vol 1. Section 10, Steam and power conversion system, through section 22, 'Conclusions'. NUREG-1503. Washington, DC
Scanlon TM (1998) What we owe to each other. Belknap Press of Harvard University Press, Cambridge, MA
Schulz T (2006) Westinghouse AP1000 advanced passive plant. Nucl Eng Des 236(14–16): 1547–1557
Seaborg GT (1962) The first nuclear reactor, the production of plutonium and its chemical extraction. IAEA Bull 4:15–17
Silady FA, Cunliffe JC, Walker LP (1991) The licensing experience of the modular hight-temperature gas-cooled reactor (MHTGR). Energy 16(1–2):417–424
Spiewak I, Weinberg AM (1985) Inherently safe reactors. Annu Rev Energy 10(1):431–462
Taebi B (2011) The morally desirable option for nuclear power production. Philos Technol 24(2):169–192
Taebi B (2012) Intergenerational risks of nuclear energy. In: Roeser S, Hillerbrand R, Sandin P, Peterson M (eds) Handbook of risk theory. Epistemology, decision theory, ethics and social implications of risk. Springer, Dordrecht, pp 295–318
Taebi B, Kadak AC (2010) Intergenerational considerations affecting the future of nuclear power: equity as a framework for assessing fuel cycles. Risk Anal 30(9):1341–1362
Taebi B, Kloosterman JL (2008) To recycle or not to recycle? An intergenerational approach to nuclear fuel cycles. Sci Eng Ethics 14(2):177–200
Taebi B, Roeser S, Van de Poel I (2012) The ethics of nuclear power: social experiments, intergenerational justice, and emotions. Energy Policy (51):202–206
Tester JW, Drake EM, Driscoll MJ, Golay MW, Peters WA (2005) Sustainable energy: choosing among options. MIT Press, Cambridge, MA
Valentin J (2013) Radiation risk and the ICRP. In: Oughton D, Hansson SO (eds) Social and ethical aspects of radiation risk management. Elsevier, Amsterdam, pp 17–32
Van de Poel IR (1998) Changing technologies, a comparative study of eight processes of transformation of technological regimes. PhD dissertation, University of Twente, Enschede
Van de Poel IR (2009) Values in engineering design. In: Meijer A (ed) Philosophy of technology and engineering sciences. Elsevier, Amsterdam, pp 973–1006
Van de Poel IR, Royakkers LMM (2011) Ethics, technology and engineering. An Introduction. Wiley-Blackwell, West Sussex
WCED (1987) In: Brundtland GH, Angelli S, Al-Athel S, Chidzero B (eds) Our common future. World Commission on Environment and Development (WCED), Oxford
Weinberg AM, Spiewak I (1984) Inherently safe reactors and a second nuclear era. Science 224:1398–1402

# Design for Values in Software Development

Huib Aldewereld, Virginia Dignum, and Yao-hua Tan

## Contents



**Abstract**

A rising trend in software development by public and private organizations is the focus on solutions that combine services (potentially provided by others) into value-added systems. ICT systems based on service-oriented and distributed computing principles require profound changes in the way the software is designed, deployed, and managed. The software that organizations develop and/or use needs to comply with all sorts of values, e.g., legal norms (privacy) or societal values (be environmental friendly), yet no software development methodology currently handles values in the design process (explicitly). Existing "waterfall"-like software development falls short in this new multidimensional field, where approaches are required that can integrate values, functionalities, and behaviors in system design. In this chapter, we introduce the Value-Sensitive Software Development (VSSD) framework as a Design for Values approach to the development of ICT systems. VSSD aims to make the

H. Aldewereld (✉) • V. Dignum • Y.-h. Tan
Delft University of Technology, Delft, The Netherlands
e-mail: h.m.aldewereld@tudelft.nl; m.v.dignum@tudelft.nl; y.tan@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_26

relations between the values, the domain, and the software product explicit, thereby improving maintainability of the software.

**Keywords**
Software development • Service-oriented architectures • Norms • Institutions

## Introduction

Suppose you are a web designer creating a new web page for the Delft University of Technology to offer massive open online courses (say, on Design for Values) to interested people. In the analysis phase, as is typically done in software development, the requirements of the various stakeholders (e.g., Delft University of Technology, students, lecturers, etc.) are investigated and formulated as functional and non-functional requirements.[1]

One of the functional requirements for this web page may be to ask the user permission to use cookies to provide a better, more personalized browsing experience, and this requirement may be motivated by the value of guaranteeing the privacy of the students (as per EU Directive 2002/58/EC). If, however, in a later phase it is decided to build an application for smartphones and tablets instead, this requirement is discarded, because such devices do not use cookies. Yet by discarding this requirement, the underlying reason to ask for the use of cookies, to guarantee user privacy, is also lost.

This example shows, in simple terms, the problem in current software development approaches; values are currently only implicitly involved in the development process. The link between the values and the application is, at best, implicit in the decisions made and the choices taken. In the requirements elicitation process (Sommerville and Sawyer 1997), the values are translated to concrete requirements. These requirements are then used throughout the development process, and the related value is lost. However, due to their abstract nature, values can be translated to design requirements in more than one way. If the values and their translation to requirements are left implicit in the development process, one loses the flexibility of using alternative translations of those values.

In this chapter, we will show how to describe the links between values, requirements, and software design. Making these links explicit allows for improvements in the traceability of (the effects of) the values throughout the development process. This traceability greatly increases the maintainability of the application. That is, if a value is to be implemented differently, the explicit links between the values and the application make it much easier to determine which parts of the application should be updated.

[1]Functional requirements specify required software behavior (e.g., registration should happen before giving users access to the course); non-functional requirements specify quality aspects of the application's operation (the application should be usable by all users).

Such a Design for Values approach to the development of ICT systems brings about several challenges for how applications should be designed, managed, and deployed. In particular, development methodologies and software engineering paradigms must be able to analyze and incorporate values into the design and implementation. This means that application development processes should enable keeping track of the underlying values, requirements, and objectives that motivate the interaction among the different peers. Design for Values approaches, which make explicit the values behind implementation decisions, will need to answer the following questions:

- How to elicit, maintain, and integrate the values and requirements of each stakeholder?
- How to develop applications without compromising the values and motivations of the stakeholders involved?
- How to manage application execution when that application requires interactions between various software components, especially when these components are built or maintained by different organizations, with potentially different values?

These issues point to the need of a multilayered approach to software development where the links to the values are kept explicit. In this chapter, we present the Value-Sensitive Software Development (VSSD) framework that takes an integrated view of software applications, combining the values and global objectives of the application with existing structures and the dynamics of the application and of its context.

This chapter is organized as follows. The next section gives a further motivation for this work and presents a background in Design for Values and service-oriented architectures. In the third section, we describe the VSSD framework that enables linking values and operational concerns of applications. A case study is presented in the fourth section to illustrate the use of VSSD. Finally, we end this chapter with a discussion of open issues and future work in the fifth section and conclusions in the sixth section.

## Motivation and Background

During software development, many choices need to be made at a high level of detail, which shape the nature of the resulting application. The reasoning underlying each decision is ultimately based on abstract organizational and (individual) stakeholders' values like integrity, trust, security, or fairness.

A failure to comply with such values can lead to the resistance to the introduction of ICT solutions by the organizations or the society (van den Hoven 2007). There are many examples of such ICT failures in the past (Prins 2011); some of the prime examples in the Netherlands are the electronic patient dossier (EPD)[2],

[2]The Electronic Patient Dossier was meant to be a national system to collect medical information of all patients, but failed due to a lack of privacy.

the OV-chip card[3], and C2000.[4] A consequence is that each of these ICT projects has their development running over budget. The need for applying Design for Values approaches in ICT seems necessary now more than ever. In the "Design for Values" section, we provide a background on Design for Values.

Given the complexity of ICT systems nowadays, software development is no longer focused on the creation of large monolithic applications that can solve everything. A rising trend in software development is to create smaller, simpler, and more specialized pieces of software, called services, and combining these services into interaction patterns to solve the complex problems. This idea of software development through the combining of services is called service-oriented architectures (SOAs). In the "Service-Oriented Architecture" section, we provide a background on SOA approaches.

## Design for Values

The recognition that values have an impact on the development of technology comes from the field of philosophy (ethics, in particular). Especially the topic of Design for Values has seen much development recently.

Design for Values is a methodological design approach that aims at making moral values part of technological design, research, and development (van den Hoven 2005). Values are typically high-level abstract concepts that are difficult to incorporate in software design. In order to design systems that are able to deal with moral values, they must be operationalized while maintaining the traceability of its originating values. The Design for Values process aims to trace the influence of values in the design and engineering of systems.

In order to take a Design for Values perspective on the development of ICT systems, it is necessary to "open the black box of technology design and development and describe its rich and heterogeneous content, and make an inventory of the degrees of freedom in the design and engineering process, which choices have been made and can be made. . ." (van den Hoven 2007). This requires the use of methodologies that satisfy the following principles: (1) global aims and policies should be explicitly described; (2) enforcement is context based and should be negotiated between stakeholders; and (3) design decisions should be formulated explicitly rather than being implicit in the procedures and objects.

[3]The OV-Chip card was introduced as an universal, electronic card for all public transport in the Netherlands to replace the various different tickets in use; the introduction was met with resistance due to shortcomings of the security of the card.

[4]C2000 is a shared, private, communications band for all crisis management forces (police, medics, fire brigades), which failed due to insufficient consideration of the procedures/policies of the involved stakeholders.

In particular, value descriptions are usually done at an abstract level and do not provide enough formality to be usable at the system specification level. Therefore, the first step in Design for Values is to provide a formal representation of values that "translates" natural language description into formal values in a formal language.[5] The translation to formal values will provide the basis for the remainder of the Design for Values process, eventually leading to a system that supports agents in direct control and self-regulative contexts. However, the relation between abstract values and formal norms is more complex than mere formal interpretation. In society, institutions are defined as "the set of rules actually used by a set of individuals to organize repetitive activities that produce outcomes affecting those individuals and potentially affecting others" (Ostrom et al. 1994). Institutions set the necessary preconditions for individual interactions (Scharpf 1997) and as such provide structured interpretations of the concepts in which norms are stated. In particular, institutions do not only consist of norms but also describe the ontology of the to-be-regulated domain. For instance, whether something within a given institution counts as personal data and should be treated as such depends on how that institution interprets the term personal data (Vázquez-Salceda et al. 2008). This perspective on institutions, which emphasizes the semantic dependencies between abstract and concrete norms, goes hand in hand with acknowledged positions in the study of social reality and legal systems, such as the concept of "counts as" (Searle 1995; Jones and Sergot 1993). Grossi et al. (2005, 2006) and Aldewereld et al. (2010) proposed, investigated, and applied a framework for formally representing such statements, where the relation between X and Y is interpreted as a standard concept subsumption, but which holds only in relation to a specific context.

Finally, value has also been given focus from a business perspective. Most notable of these are the development of the Resource-Event-Agent (REA) and E3Value ontologies (McCarthy 1982; Hrubỳ et al. 2006; Gordijn et al. 2001). REA originated from traditional business accounting as a basis for accounting information systems. E3Value was created to identify exchanges of resources (hence value) between actors in a business case. Both ontologies include concepts on operational and knowledge levels. These ontologies are rather successful for modeling business domains and are closely tied with implementation frameworks (Gordijn et al. 2006). However, the focus of these ontologies lies on economic value. Their level of detail in modeling the domain is not abstract enough to also include the philosophical/ethical types of values. Due to their wide use for business modeling (and close relation to service-oriented architecture implementations), REA or E3Value can be seen as a possible bridge between Design for Values and service-oriented architectures.

[5]For details on the formalization of values, we refer interested readers to (Aldewereld 2007; Vázquez-Salceda 2004).

## Service-Oriented Architecture

Service-oriented architecture (SOA) is an IT architectural style that supports the vision of systems as a set of linked services, or repeatable business tasks, that can be accessed when needed over a network or the Internet (Erl 2005). These services can coalesce to accomplish specific business tasks, enabling organizations to quickly adapt to changing conditions and requirements. The service-oriented architecture paradigm is grounded on the idea that ICT system design and implementation is guided by strategic business goals, to ensure the positive transformation of organizations and realize their missions (Versteeg and Bouwman 2006). Design approaches to service-oriented architectures extend traditional software development methodologies and architectures to support the development life cycle of service-based systems, starting at the conceptualization phase, supporting design and architecture activities, and extending modeling best practices for service operations in a production environment (Papazoglou and van den Heuvel 2007).

The deployment of systems based on service-oriented architectures is becoming widespread and successful in many application domains. Numerous commercial and civil organizations are actively engaged in converting their existing information systems or constructing new systems based on SOA principles (The Open Group 2008). Nevertheless, service-oriented analysis and design methods mostly concentrate on the operational aspects of modeling and are not able to capture organizational values, vision, or context. This means that decisions, choices, and their context are not explicit such that they can guide later analysis and transformations. That is, the successful deployment of SOAs requires dedicated methodological approaches that address the specific aspects of the service paradigm. Moreover, current SOA-based systems are *static*, in that design requires services and their possible interactions to be defined and characterized in advance. The use of *dynamically* discovered, configured, deployed, engaged, and maintained services has not been successful yet (Bell 2008). The problem here is that current service standards, which are necessary for widespread usage of services, only allow for the specification of simple syntax and formatting of service invocations (Papazoglou and van den Heuvel 2007).

The features of service governance are well beyond what was originally envisioned for SOAs. The features include quality of service (QoS) and contracts, i.e., service-level agreements (SLAs) among the participants. Moreover, to make this governance dynamically and autonomously configurable, the participants will need the ability to negotiate at run time to establish the SLAs, to monitor compliance with them, and to take actions to maintain them. These are capabilities associated with *software agents* which have been advocated as the only reasonable solution for the problems described above (Brazier et al. 2012). Software agents are encapsulated computer systems that are situated in an environment and that are capable of flexible, autonomous action in that environment to meet their design objectives (Woolridge and Jennings 1995). The introduction of agents can in fact increase the flexibility of service interactions (Huhns 2002; Payne 2008).

Even if the autonomy of agent-based services can make them less predictable, autonomy also enables them to recover from failure and to avoid deadlocks and livelocks, thereby making them more reliable. Their interactive nature can improve the efficiency of agent-based services and can increase the flexibility of service interactions, but at the cost of concurrency problems. More importantly, the ability of agent-based services to negotiate and reconcile semantic differences makes an agent-based approach very suitable to integrate the different levels of abstraction and scoping that are associated with the incorporation and operationalization of values in the design of service-based IT systems.

## The Value-Sensitive Software Development Framework

Architectural decisions capture key software design issues and the rationale behind chosen solutions. They are conscious development decisions concerning a software application as a whole, or one or more of its core components, with impact on nonfunctional characteristics such as software quality attributes. An increasing number of SOA methods are promoting the idea that these architectural decisions should be made explicit. The decisions are grounded on both domain characteristics and the values of the designers and other stakeholders involved in the development process. Taking a VSD perspective, it becomes possible to make explicit the link to the values behind architectural decisions.

The VSSD (Value-Sensitive Software Development) framework presented in this section is one such method to make the links between the values (as the motivation for the architectural decisions), the domain characteristics (as shaping the architectural possibilities), and the resulting software. VSSD presents a basis to specify and analyze complex adaptive software from a value-sensitive perspective. VSSD is based on the OMNI framework (Vázquez-Salceda et al. 2005) for complex distributed organizations. The VSSD conceptual framework can be considered from three complementary perspectives or views:

- The *values view* shows the operationalization process of stakeholders' mission (or power of intention) as high-level mechanisms to protect essential (human) values and as such putting forward vision on their own responsibilities and aims. This process specifies mechanisms of social order, in terms of values, norms, and procedures. This view relates most with a philosophical perspective on value-sensitive design (see, e.g., van de Poel (2013)).
- The *modeling view* exposes the architectural design of the system, based on value decisions and domain characteristics, specifying system characteristics at increasing level of detail. This view relates most to existing service-oriented architecture (SOA) design methodologies. SOA has an existing abstraction layering between the system specification (choreography) and system enactment (orchestration). These abstraction layers relate nicely with the aim of VSSD to make decisions at each level explicit.

- The *business view* exposes the contextual aspects of the domain, based on the existing systems and services. This view, as the name suggests, relates most with a business approach to value (McCarthy 1982; Gordijn et al. 2006).

All three dimensions can be considered at three different abstraction levels. This facilitates the analysis and design of applications, by providing the means to link abstract concepts to its (concrete) implementation counterparts. The distinction between the levels of abstraction is necessary to capture all elements at their *native* level of abstraction and to make the necessary translations between the different elements explicit. These abstractions levels are:

- The **abstract level:** where the statutes of the system to be modeled are defined at a high level of abstraction. This level can be used to model elements from a first step in the requirement analysis. It also contains the definition of terms that are generic for any system organization (i.e., that are not contextual).
- The **concrete level:** where specific model components are specified, based on the domain analysis and value design processes. The meaning of the abstractions defined at the higher level is refined in terms of concrete functionality and regulatory concepts and their interdependencies.
- The **implementation level:** represents the implementation phase of the development process. This level assumes given system components (e.g., services) as basis for the implementation, including mechanisms for role enactment and norm enforcement.

These dimensions and levels present the means to explicitly represent the relations between values, domain, and software (see Fig. 2). The motivation for the architectural decisions (the *why*), describing the reasons why things are as they are, is captured in the values view. The values, on the abstract level, and the norms and procedures on the more concrete levels represent the incentives for the way the software is as it is. The software is not designed just because of the values, however, but is also shaped by domain characteristics, which are represented on the business view of VSSD. The business view, in this sense, represents the *what*, that is, the driving forces from business and domain perspectives (e.g., existing infrastructure, available services, and functional requirements). Finally, the modeling view of VSSD brings these two forces together in the software design. It describes the *how* of the software, meaning the way in which the domain-specific constraints and the value-based incentives are used to shape the software.

Furthermore, we can identify three core activities in the VSSD framework. First, there is the *elicitation*, which is located between the values view and the modeling view. This activity is concerned with finding the applicable values and the way these influence the application design. Second is the *development*, which is located between the business view and the modeling view. Motivated by the values (and hence, the elicitation), the model of the software is developed while taking into account the domain constraints. Finally, the *execution* activity is the result of the modeling, which happens when the software is deployed.

## Application

Let us return to our example from the introduction. As mentioned, we are interested in designing a web page for the Delft University of Technology to offer massive open online courses (MOOCs). From the requirements elicitation, we got the (functional) requirement that we request users the permission to store information in a cookie. The cookie is required to provide a more user-friendly browsing experience; for example, the site could remember which courses you already followed. Let us assume that there is another requirement to provide support for impaired people; we could, for instance, include a text-to-speech functionality to read out loud parts of the website to people with visual impairments.

There are two hidden values at stake here: privacy and equality. The former, because the data we store in the cookies is of a personal nature, and consent has to be given before storing such data. The latter, because the inclusion of a text-to-speech functionality to assist impaired users is based on the idea that everyone should be able to use the website. Storing personal data while ensuring privacy also requires the upholding of another value: security. The relations between the software, the requirements, and the underlying values are represented in Fig. 3 above.

The values on the left-hand side of Fig. 3 represent the motivations and underlying values of the application. The requirements, and domain restrictions, are represented on the right-hand side of Fig. 3; they represent the manner in which the domain restricts the application design, which existing infrastructures (services) are available, and in what way the objectives should be achieved. The middle part of Fig. 3 represents the implementation choices that end up in the runtime of the application. These are the characteristics of the software, describing concrete execution features of the application.

The strength of VSSD is in the explicit representation of the links between these layers (in Fig. 3, only shown explicitly in the vertical dimension of the values view). Keeping explicit the motivations for particular design choices, either in the execution or in the requirements, means alternative implementations can easily be explored without the need to redesign the application ground-up. The links explain the reasons why particular execution elements exist (because they implement a particular value) and/or why a particular software behavior was formulated (because that particular procedure was required by a high-level norm or value). When changing the application, e.g., from a web page to a tablet app, these explicit links to the motivations can help you find the appropriate redesign steps. For example, in the redesign, instead of asking for cookies (which are not used on tablet apps), you could enable a secure connection for registration and then perform all communication between the client (the tablet) and the server in an anonymized manner. Since the data stays on the tablet, and one can assume that the user owns that device (and is thus responsible for its security), these choices also implement the required value of privacy.

Since the above example is rather simple, the relation between the requirements and (hidden) values is quite straightforward. For more complex applications, with many more requirements, these relations might not be so obvious, and explicit links

between the values, requirements, and execution are direly needed to ensure the value sensitivity of the application.

## Open Issues and Future Work

The explicit use of values in software development improves the traceability of the effect(s) of the values throughout the development process. This increase in traceability allows for shorter (and thus more flexible) development cycles, which in turn allow for iterative development (i.e., adaptive development instead of progressive development). Moreover, the traceability greatly increases the maintainability of the product. The part(s) of the system that need to be updated due to a change in a value are much easier to determine due to the explicit links between the values and the implementation.

All of these benefits follow from the assumption that the conceptual framework of VSSD, as presented in Fig. 1, can be transformed into a software development methodology. The creation of a value-based system software development methodology requires the existence of:

(a) a theoretical model
(b) design procedures to guide the process

The theoretical model is presented in this chapter (see Fig. 1). The design procedures, however, are still lacking. While the vertical translations (within each view of VSSD) are well understood, the relations between the different views are still largely missing. For the values view, Aldewereld (2007) and van de Poel (2013) give formal relations between the values, norms, and procedures. The E3Value ontology (Gordijn et al. 2006) contains the elements and relations needed for modeling the business view. The SOA paradigm (Arsanjani 2004) covers most of the modeling view.

In future studies, we intend to investigate the relations between the views of VSSD. An important set of these relations are the ones that deal with the upper abstraction level of VSSD. To correctly introduce values in the design, they need to be captured at their original level of abstraction. But how does one systematically combine the values with a (abstract) description of the domain to structure the motivations of the system? That is, how does one combine the abstract values and abstract business concepts to come to an abstract system design (represented as the service organization in the abstract modeling view of VSSD)? Moreover, when such an abstract design is possible, how does one connect this to the SOA design paradigm (which represents the concrete and implementation levels of the modeling view) to create systems that comply with the values *by design*?

The success of a software development process that uses values explicitly as proposed above boosts the importance of another field of research that has, until recently, gained little interest from businesses: the elicitation of values. The elicitation of values has been explored in, for instance, the field of ethics

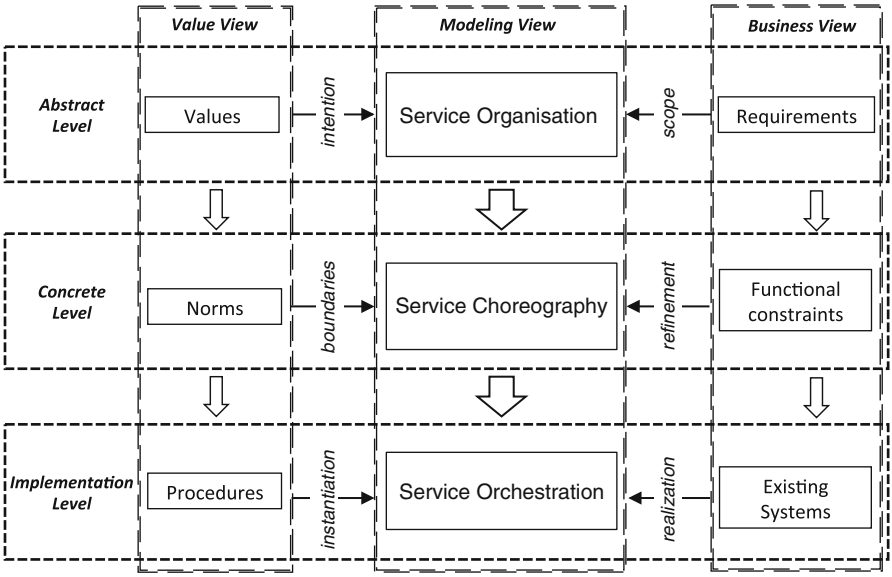



**Fig. 1** Levels and views in the VSSD framework

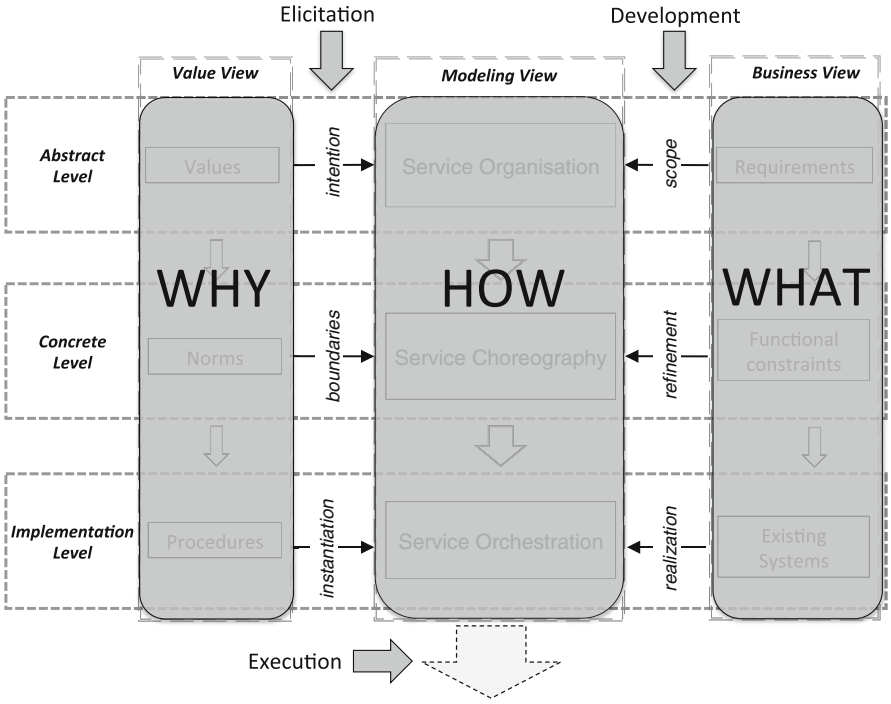



**Fig. 2** Aspects and activities in VSSD

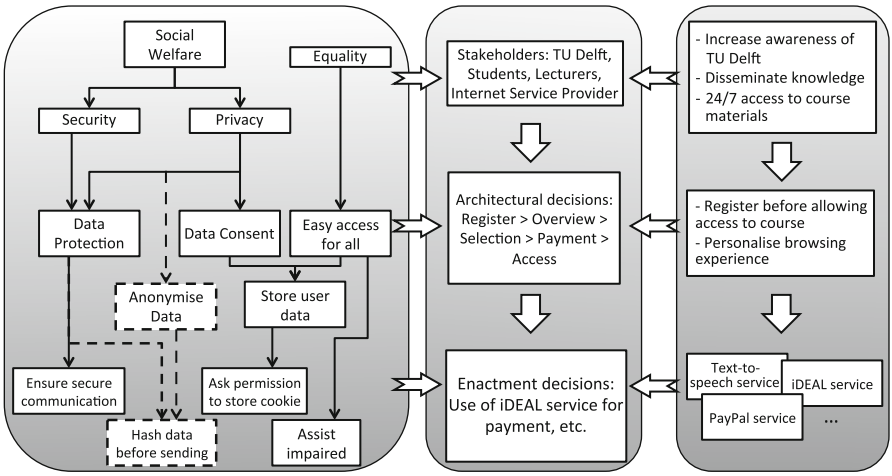



**Fig. 3** VSSD applied to MOOC-page design

(e.g., see Pommeranz et al. (2012)) but has not yet been utilized in design processes. While designing for values is of big importance to gain software that is compliant by design, the elicitation of the values is a big success factor in that process. If values are elicited incorrectly or incompletely, the resulting design of the software may be faulty as well.

## Conclusion

ICT systems based on service-oriented and distributed computing principles require profound changes in the way software systems are designed, deployed, and managed. In particular, we identify a need to address the consequences of changing environments, on-demand interactions, and increasing demands on security and safety, for the design and deployment of such software systems. Existing "waterfall"-like engineering techniques, where software is built in a single continuous process from requirements to product, fall short in this new multidimensional field, where approaches are required that can integrate values, functionalities, and behaviors into running systems and support active, distributed, independent processes.

In the introduction, we mentioned three questions that a Design for Values approach should be able to answer:

1. How to elicit, maintain, and integrate the values and requirements of each stakeholder?
2. How to develop applications without compromising the values and motivations of the stakeholders involved?

3. How to manage application execution when that application requires interactions between various software components, especially when these components are built or maintained by different organizations, with potentially different values?

In this chapter, we presented the Value-Sensitive Software Development (VSSD) framework as a Design for Values approach to the development of ICT systems to answer the first and last of these questions (questions 1 and 3). The VSSD framework combines existing work on the modeling of service-oriented computing (modeling view) with modeling techniques in the description of the values of the system (values view) and the intended application domain (business view).

Orthogonal to the perspectives on the modeling of software systems, the VSSD frameworks explore those views on all possible levels of abstraction. Making an explicit distinction on the level of abstraction makes it easier to understand in which way (and at which stage of the development) elements in the values and business view influence system modeling. Moreover, we argued that maintaining explicit links between the abstract and concrete model elements allows for easier maintenance of the system, a crucial element missing in current methodologies, but direly needed in the dynamic fields of, for example, service-oriented approaches, systems of systems, and Web 3.0 (Internet of Things).

The current chapter, being focused on Design for Values, highlighted the value view of the VSSD framework. We explored how values are to be used in the development of dynamic interconnected systems. In particular, we looked into how the abstract values can be translated to the more concrete elements of the framework. Values are often expressed on a high level of abstraction to ensure their longevity. This high level of abstraction, however, makes them particularly difficult to implement in (software) systems, as systems need the concrete elements for the creation of algorithms and protocols.

The conceptual framework presented in this chapter presents the first step into the explicit use of values in the design and implementation of systems. The next steps include the investigation of the relations between the views of the VSSD framework to create design procedures to guide the implementation. Moreover, with the creation of a design methodology that uses values explicitly, a new emphasis is put on the (correct) elicitation of the business values. The creation of such a design methodology will solve our last question (question 2, above) for the development of a Design for Values approach to software development. We will explore these final aspects in future work.

## Cross-References

► Design for Values in ICT
► Design for Values and the Definition, Specification, and Operationalization of Values

## References


Aldewereld H (2007) Autonomy vs. conformity: an institutional perspective on norms and protocols. PhD thesis, Universiteit Utrecht

Aldewereld H, Álvarez-Napagao S, Dignum F, Vázquez-Salceda J (2010) Making norms concrete. In: Proceedings of the 9th international conference on autonomous agents and multiagent systems, AAMAS'10, ACM Press, Toronto, vol 1, pp 807–814

Arsanjani A (2004) Service-oriented modeling and architecture. http://www.ibm.com/developerworks/library/ws-soa-design1/

Bell AE (2008) From the front lines: DOA with SOA. Commun ACM 51(10):27–28

Brazier F, Dignum V, Huhns MN, Derksen C, Dignum F, Lessner T, Padget J, Quillinan T, Singh MP (2012) Agent-based organisational governance of services. Multiagent Grid Syst 8(1):3–18

Erl T (2005) Service-oriented architecture: concepts, technology, and design. Prentice Hall PTR, Upper Saddle River

Gordijn J, Akkermans H, Van Vliet J (2001) Designing and evaluating e-business models. IEEE Intell Syst 16(4):11–17

Gordijn J, Yu E, van der Raadt B (2006) E-service design using i* and e3value modeling. IEEE Softw 23(3):26–33

Grossi D, Dignum F, Meyer JJC (2005) Context in categorization. In: Proceedings of CRR'05, workshop on context representation and reasoning, CUER workshop proceedings, In conjunction with CONTEXT 2005, Paris, France, vol 136

Grossi D, Meyer JJC, Dignum F (2006) Counts-as. classification or constitution? an answer using modal logic. In: Goble L, Meyer JJC (eds) Proceedings of the eight international workshop on Deontic logic in computer science (DEON'06). Springer Utrecht, The Netherlands

Hrubý P, Kiehn J, Scheller CV (2006) Model-driven design using business patterns. Springer, New York

Huhns MN (2002) Agents as web services. IEEE Internet Comput 6(4):93–95

Jones A, Sergot M (1993) On the characterization of law and computer systems. In: Deontic logic in computer science. Wiley, Chichester, pp 275–307

McCarthy W (1982) The REA accounting model: a generalized framework for accounting systems in a shared data environment. Account Rev 42(3):554–578

Ostrom E, Gardner R, Walker J (1994) Rules, games, and common-pool resources. University of Michigan Press, Ann Arbor

Papazoglou M, van den Heuvel W-J (2007) Business process development life cycle methodology. Commun ACM 50(10):79–85

Payne TR (2008) Web services from an agent perspective. IEEE Intell Syst 23(2):12–14

Pommeranz A, Detweiler C, Wiggers P, Jonker C (2012) Elicitation of situated values: need for tools to help stakeholders and designers to reflect and communicate. Ethics Inf Technol 14(4):285–303

Prins C (2011) iOverheid. Amsterdam University Press, Amsterdam

Scharpf F (1997) Games real actors play: actor-centered institutionalism in policy research. Westview Press, Boulder

Searle J (1995) The construction of social reality. Free Press, New York

Sommerville I, Sawyer P (1997) Requirements engineering: a good practice guide. Wiley, Chichester

The Open Group (2008) SOA case studies, SOA Source Book: https://www2.opengroup.org/ogsys/catalog/G102

Van de Poel I (2013) Translating values into design requirements. In: Philosophy and Engineering: Reflections on Practice, Principles and Process, Springer, pp. 253–266

van den Hoven J (2005) Design for values and values for design. Inf Age J Aust Comput Soc 7(2):4–7

van den Hoven J (2007) Ict and value sensitive design. In: Goujon P, Lavelle S, Duquenoy P, Kimppa K, Laurent V (eds) The information society: innovation, legitimacy, ethics and democracy in honor of professor Jacques Berleur s.j., IFIP international federation for information processing, vol 233. Springer, Boston, pp 67–72

Vázquez-Salceda J (2004) The role of norms and electronic institutions in multi-agent systems. The HARMONIA framework, Whitestein Series in Software Agent Technology. Birkhäuser Verlag, Basel

Vázquez-Salceda J, Dignum V, Dignum F (2005) Organizing multiagent systems. JAAMAS 11(3):307–360

Vázquez-Salceda J, Aldewereld H, Grossi D, Dignum F (2008) From human regulations to regulated software agents' behaviour. J Artif Intell Law 16:73–87

Versteeg G, Bouwman H (2006) Business architecture: a new paradigm to relate business strategy to ict. Inf Syst Front 8:91–102

Woolridge M, Jennings N (1995) Intelligent agents: theory and practice. Knowl Eng Rev 10(2):115–152

# Design for Values in Water Management

Wim Ravesteijn and Otto Kroesen

## Contents



**Abstract**

In view of current massive water quantity and quality problems as well as shifting social wishes and requirements, the currently dominant water engineering and management approaches need to change. The water domain is indeed the scene of reorientation and transformation, though from a perspective of design for values much remains to be desired. This chapter discusses these matters, investigates present water approaches, and indicates how value considerations have always been implicitly present in such approaches. Three approaches for dealing with water affairs are distinguished: the technical-economic approach, integrated water resources management, and the negotiated approach. In this sequence, value considerations have become increasingly important. In addition

W. Ravesteijn (✉) • O. Kroesen
TU Delft, Delft, The Netherlands
e-mail: w.ravesteijn@tudelft.nl; j.o.kroesen@tudelft.nl



J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0_31

to suggesting improvements, we will present a step-by-step plan for a more explicit design for values approach, as the way to move forward in dealing with global water issues.

**Keywords**
Water problems • Water engineering and management approaches • Revaluing • Step-by-step method

## Introduction

The pressure on water resources is increasing in the world, and, consequently, overall accessibility to and a fair distribution of water resources have a high priority on the agenda of water developers, managers, and policy makers. Other water problems draw attention as well, especially growing contamination and the increasing risk of flooding. All these problems require serious changes in water resource management and development, including an increase in human and institutional capability to deal with competing and conflicting values related to the multitude of water uses and problems (Hoekstra and Huynen 2002). We can specify the main problems as follows:

1. Clean drinking water is increasingly scarce because population growth and economic development raise the demand for water. Groundwater depletion is a growing problem. In general, water scarcity gives rise to national and international conflicts.
2. Contamination of water resources is increasing, due to increasing food and other agricultural production, population growth, and industrialization.
3. More and more people are facing large-scale flooding. River basins are progressively deteriorating as a consequence of logging as well as unsustainable cultivation methods. Rising seawater levels, melting glaciers, and more extreme weather as a consequence of global warming aggravate flooding in river basins and coastal areas.
4. Too much water through rainfall and too little water for irrigation give rise to tensions between rural areas and cities.

These water problems can be framed and tackled in a variety of ways. Roughly, in this paper we distinguish (and later specify) three of them. Water problems could be seen as merely *technical* problems, to be solved by hydraulic civil engineers (see, e.g., Ricketts et al. 2004). They could also be considered in terms of *management* issues in relation to social interests, which should be tackled by water managers and policy engineers besides civil engineers (as elaborated by, e.g., Hermans 2005). A third way and new of viewing and dealing with water problems emphasizes *social* interests and viewpoints and puts the water users and stakeholders in the leading role (as explored in, e.g., Both Ends and Gomukh 2005). Though there are major regime shifts involved in the development of these three approaches, they nowadays coexist. All three are used in designing works and

procedures for (multiple) water values, though not always explicitly. Our goal is to investigate how values are taken into account and how value considerations can be made more explicit, so improving the social and moral embedding of water technology and organization.

This chapter discusses the three approaches from a value perspective and indicates how a design for values perspective could be developed in relation to these approaches. On the one hand, it investigates and assesses what their capacities and potentials are in addressing present-day water problems. On the other hand, it explores and shows how these approaches could be upgraded ("revalued") by focusing on the often implicit underlying values and by dealing more explicitly with divergent values through negotiation and dialogue about value priorities, institutional design, and social experimentation in order to find the right way forward. In this way the authors will establish the impact of a design for values perspective in dealing with water problems. Exemplary cases will be used to illustrate the diverse water approaches as well as the usefulness of a design for values perspective and the shape it could take. Design for values is new, and this paper also aims to contribute to developing it as a clear and practical method in framing as well as understanding water-related issues.

We start with briefly introducing civil engineering and its values, followed by outlining the three dominant approaches in water engineering and management. Next, we focus on design for values, in terms of which we revisit the three approaches and discuss ways to take them to a new level of richness and applicability. Our final considerations will contain a method or step-by-step plan for follow-up work in the new domain of design for values. Throughout the paper, we adopt a systems approach that considers the value-driven design of integrated water management systems that include artifacts like barriers, dams, water treatment plants, pumps, etc., but we do not discuss these artifacts separately.

## Water Interference and the Values It Serves

Hydraulic civil engineering and water management are the prime disciplines that deal with the challenges mentioned in the introduction (Civil Engineering 2013; Cech 2009). Both disciplines have a long history, going back to the dawn of civilization, but as a science civil engineering developed in modern history, while water management as a field of activity in its own right can be seen as a postwar offshoot. During the nineteenth century civil engineering professionalized along with the training of engineers and the application of science, including mathematics, as hallmarks. Generally, it comprises the construction of water works (dams, barriers, sluices, and other structures) and systems (like irrigation, flood defense, and drink water systems). Operation and maintenance introduced scientific water management, which more and more became a framework for building and even an approach in itself (the nonstructural approach, see below). Nowadays, water resources development and management is as much a topic of experts as it is a domain with input from stakeholders and society in general.

Water is essential to life and it serves many purposes. Water resources development and management are crucial to human society, the health and welfare of the people, and processes of socioeconomic development and transformation (Sachs 2008). Consequently, a range of human values is involved in water interference, while a say of "stakeholders" is almost inevitable. Historically and presently, water efforts take place from a variety of values, implicitly or explicitly. Key values include *safety* (against flooding), *security* (food and drinking water), *utility* (cooling, industrial water, waste, shipping, land acclamation, energy), *sustainability* (ensuring quantity and quality). Focused on management, an additional key value is *distributive justice* (equal access to common goods) and along with that *social sustainability* (addressing the needs of the poor and fighting inequalities in the world). Other values are historically and culturally determined, like democracy (as in the case of the Dutch water associations), although in a broad sense, in view of the many stakeholders involved, many issues may not be resolved without some form of *democratic participation*. Often the utility priorities of water purposes are historically and culturally determined (see, e.g., Dubbelman 1999; cf. Song et al. 2011). These are all key values to be considered in integral water management systems. Often in their tensions and trade-offs, they reflect the value priorities of the different stakeholders involved, and for that reason they cannot be resolved without negotiation or dialogue. Because water resources are limited, and many values are involved in water use, different socio-technical regimes have been developed to steer and balance water works construction and management, ranging from simple water rights to elaborate systems of law and regulation (see, e.g., Hoekstra and Huynen 2002; Kissling-Näf and Kuks 2004).

## Current Water Engineering and Management Approaches

Water engineering and management have become more complex over time as the number of interests and values water works had to accommodate grew. Maybe with the exception of some poor countries, this development took place all over the world (Ravesteijn et al. 2002). The water history of the Netherlands exemplifies this development. Dutch water system builders started with land reclamation and fighting flooding. In the course of time, water management and resources development was also done for the purposes of water distribution and shipping, ensuring water quality and fighting salinization, while recently ecological aims were included (Dubbelman 1999; van de Ven 2004). In the Netherlands and elsewhere (sometimes to a lesser degree), this development led to the construction of multipurpose works and integrated water management in order to accommodate diverging and competing interests (Disco and van der Vleuten 2002; Kissling-Näf and Kuks 2004). At the same time a process of professionalization of water engineering and management took place (Lintsen 1998) accompanied by a mix of centralization and increasing stakeholder involvement (Hermans 2005). Recent trends are fully integrated management and governance (Bressers and Kuks 2004), and a room for the water policy in combination with more attention to

ecological values (Disco 2002), both reflecting and articulating European and global tendencies (Ravesteijn and Kroesen 2007, see further below). However, more recently, a decentralized approach emerged in some areas in the world in which water users take the lead in management and engineering (Both Ends and Gomukh 2005). Consequently, we distinguish three broad approaches governing current water practices (similar to and specifying the approaches we mentioned in the introduction; cf. Both Ends and Gomukh 2005):

1. The technical-economic approach focuses on constructing water works, e.g., river engineering works like dikes, dams, and sluices; the dominant actor is the civil engineer.
2. Integrated water resources management (IWRM) focuses on managing competing or conflicting water uses through balancing the interests and values of all stakeholders; the dominant actor is the policy engineer or the process manager.
3. The negotiated approach focuses on cooperation and coordination among stakeholders through self-organization; there is no dominant actor other than the manifold users and stakeholders themselves.

These approaches have been developed sequentially in time. They reflect, in their evolution, the increasing complexity of water issues and the increasing inclusion of stakeholder dialogue, as we will show. Conflicting interests and values are reconciled and managed in different ways through these approaches. The technical-economic approach strives at controlling conflicts by constructing multipurpose works. The example here is big dams. IWRM seeks conflict control through management, with integrated river basin management (IRBM) as its model. The start of IRWM has a clear demarcation point in time and quickly became increasingly accepted. It gradually leads to more and more stakeholder dialogue, and to negotiation. The negotiated approach, which puts dialogue and value differences on the agenda deliberately and explicitly, is new and emerging, for instance, in Bangladesh. It has not yet reached wide acceptance, but according to the authors, it is promising since planning and control from above increasingly reach their limitations. These approaches as well as some of their problems are elaborated below.

## The Technical-Economic Approach: Big Dams

In present-day society, big dams are important for water supply, flood management, and electricity generation, and as such they are considered as "icons of progress." The first big dams were constructed in colonial areas (the Aswan Dam in Egypt was the very first) and in the USA (the Colorado River) and at about the same time in Russia (Dnjepr). Nowadays, big dam building especially takes place in emerging economies like China and India, with at present a true scramble for building big dams in Africa. They have proven their utility: the electrification of the railways in the USA, e.g., would not have been possible around 1900 without big water

**Table 1** Functions of big dams and examples

| |
|---|
| **Power generation**: Hydroelectric power is a major source of electricity in the world. Many countries have rivers with adequate water flow that can be dammed for power generation purposes. For example, the Itaipu on the Paraná River in South America generates 14 GW and supplied 93 % of the energy consumed by Paraguay and 20 % of that consumed by Brazil as of 2005 |
| **Water supply**: Many urban areas of the world are supplied with water abstracted from rivers pent up behind low dams or weirs. Examples include London with water from the River Thames and Chester with water taken from the River Dee. Other major sources include deep upland reservoirs contained by high dams across deep valleys such as the Claerwen series of dams and reservoirs |
| **Stabilize water flow/irrigation**: Dams are often used to control and stabilize the water flow, often for agricultural purposes and irrigation. Others such as the Berg Strait Dam can help to stabilize or restore the water levels of inland lakes and seas, in this case the Aral Sea |
| **Flood prevention**: Dams such as the Blackwater Dam of Webster, New Hampshire, and the Delta works are created with flood control in mind |
| **Land reclamation**: Dams (often called dykes or levees in this context) are used to prevent ingress of water to an area that would otherwise be submerged, allowing its reclamation for human use |
| **Water diversion**: A typically small dam used to divert water for irrigation, power generation, or other uses, with usually no other function. Occasionally, they are used to divert water to another drainage or reservoir to increase flow there and improve water use in that particular area |
| **Navigation**: Dams create deep reservoirs and can also change the flow of water downstream. This can in return affect upstream and downstream navigation by altering the river's depth. Deeper water increases or creates freedom of movement for water vessels. Large dams can serve this purpose but most often weirs and locks are used |
| **Recreation and aquatic beauty**: Dams built for any of the above purposes may find themselves displaced in course of time of their original uses. Nevertheless the local community may have come to enjoy the reservoir for recreational and aesthetic reasons. Often the reservoir will be placid and surrounded by greenery and convey to visitors a natural sense of rest and relaxation (Dam 2013; see for a Dutch pioneer in the construction of multipurpose water reservoirs Ravesteijn 2002) |

reservoirs. Global warming has put large dams even higher on the agenda than they already were, especially in developing countries: they are supposed to fit in with a low-carbon energy strategy, while they seem instrumental in adaptive water management. However, recent research has revealed that $CO_2$ emissions from organic sources like flooded vegetation and washed down detritus are gigantic from big reservoirs (International Rivers 2013), which adds up to other problems that were already known: river degradation, disappearance of wetlands, people who have to be relocated, etc. This has supported the power and influence of the international anti-dam movement, and, nowadays, more and more dams are decommissioned or plans are not executed (McCully 2003, 2011). At the same time, there is an intensified search for alternative solutions. That is not so easy, as big dams are multipurpose works. They are junctions of socio-technical systems (for energy, water, navigation, etc.), and, consequently, alternatives have to be formulated for each and every function, not only separately but also in connection with one another (see Table 1).

Indicative of the change currently taking place is a report submitted to the world commission on dams by the South Asian Network on Dams, Rivers and People (Assessment 1999), specifically commenting on India's plans for building a host of dams, especially along the Himalaya, to cover the water needs of big cities. Increasingly India's cities are supplied with water from areas further away, since the nearby sources have been depleted (groundwater) or polluted (rivers). The argument of the report goes against building big dams, because this inclination of cities to seek resources further away and to build dams to meet their requirements leads to increasing conflicts between regions and between cities and farming communities. Instead the mismanagement of water provisions in India's cities should be tackled, which in some cases leads to more than 50 % of unaccounted water.

## IWRM/IRBM: The European Framework Directive

Integrated river basin management provides a basin-wide approach, combining technological development (esp. big dams!), economic development, as well as multi-actor cooperation between the various subnational regions and countries involved (Kates and Burton 1986). IRBM is a "nonstructural" approach focused on integrated management. It is called "nonstructural" because it is less focused on building concrete structures. Thereby it overcomes the dominant technical focus of earlier periods, and thereby it also creates more room to discuss different stakeholders priorities and accordingly different value priorities. In addition, IRBM presupposes – or is greatly facilitated by – the availability of modern information and communication technologies, both in collecting information (e.g., Geo Information Systems) and in sharing it with stakeholders. It originated in the USA, where Gilbert White formulated the concept on the basis of a worldwide inventory and assessment of experiences with it. The first example was the Tennessee Valley Authority, founded in 1933 as a part of Roosevelt's New Deal Policy. In fact, IRBM being a management tool strongly reflects the doctrines of planning, which originated from the Second Industrial Revolution taking place in Germany and the USA. Planning procedures as Roosevelt's New Deal became a dominant policy instrument first in the Soviet Union. After the Second World War, planning as part of management became widely adopted. Similarly, IRBM has been broadly accepted as a framework for water policy, in which water developers often combine modern concepts with endogenous water traditions.

IRBM has spread all over the world, including Europe. In December 2000, the Water Framework Directive (WFD) was issued after long time cross-national negotiation and implemented in all 25 EU member states. The main objectives of the WFD are:

- Expanding the scope of water protection to all waters (inland surface waters, transitional water, coastal waters, and groundwater) in a holistic way
- Achieving a "good condition" for all waters by the target date of 2015, satisfying human needs, ecosystem functions, and biodiversity protection

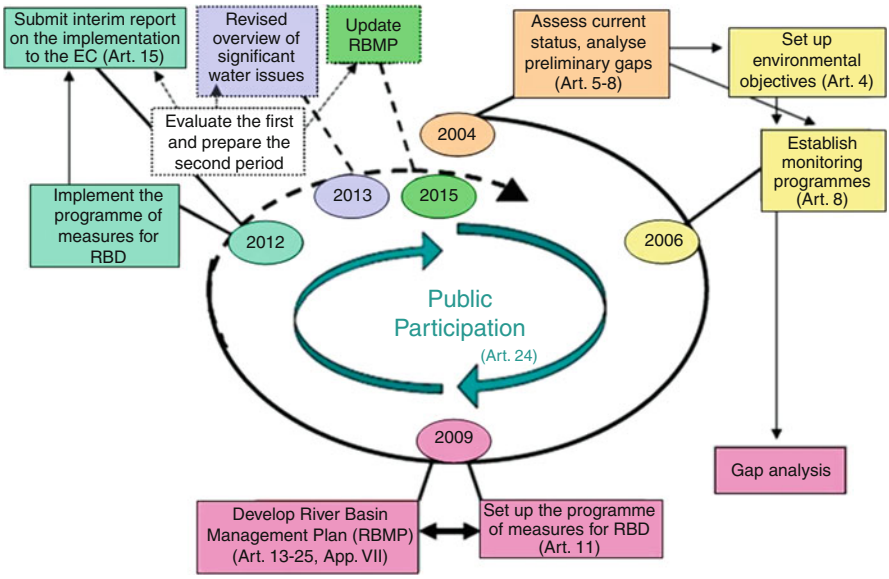



**Fig. 1** River basin management planning process (Europe Environmental Agency 2009)

- Water management within the hydrological/geographical boundaries of river basins through effective cooperation of all administrations and institutions involved
- Combined approach towards the control of both point and nonpointed pollutant sources with emission limit values and water quality standards
- Getting the right prices of water with the elements of cost recovery and cost-effectiveness provisions
- Getting citizens more closely involved in river basin management activities
- Streamlining legislation by repelling existing fragmented and burdensome regulatory systems (European Commission EC 2000)

The European WFD provides a common framework for water policy employing integrated approaches and innovative instruments to water management. It was developed as a response to the fragmented European environmental legislation. However, the international cooperation concerning the Rhine, starting in the nineteenth century, can be considered a pioneering effort. The WFD offers a policy frame for protecting and improving the quality of water resources, though it also involves flood and groundwater quality control (for which follow-up directives were issued; see below). It provides for cooperation at the level of river basins. A precise planning procedure for river basin management is part of it (see Fig. 1). The main point is towards water resources management at river basin level in districts. These river basin districts (RBDs) are largely based on surface water catchments, together with the boundaries of associated groundwater and coastal

water bodies. In April 2009, the conference of "Active Involvement in River Basin Management – Plunge into the Debate!" in Brussels has supported the preparation of river basin plans (European Water Conference EWC 2009) .

The WFD got a supplement in 2006 with the Directive on the protection of groundwater against pollution and deterioration (Resources Directive 2006). In 2007, another step was taken in realizing IRBM in Europe when the European Parliament adopted the Floods Directive (Floods directive 2007) . This requires European countries to make flood risk assessments, focused on impacts on human health and life, the environment, cultural heritage, and economic activity, and to do so before 2012. These assessments will be used to produce flood hazard and risk maps (planned to be ready by December 2013), on the basis of which flood risk management plans will be drawn up (to be ready by December 2015). The latter plans will be developed together with all relevant stakeholders and should provide policy makers and developers with a base for measures that are both technically appropriate and socially acceptable.

The EU WFD and its supplements, tailored to the specifics of the European countries, reflect many but not all the elements of IRBM. Experiences thus far are positive. Though it is causing tensions here and there in view of national water traditions (Ravesteijn and Kroesen 2007), in general, the WFD seems to work quite well. Ecological values are prominent. The goal at least – in 2015 all waters in a "good condition," that is to say, within agreed upon maximum levels of pollutants – is still within reach. However, recent assessments point out that at least 40 % of the European water bodies are "at risk" of not meeting the goal (European commission 2008). Nevertheless, the WFD has been hailed as a front-runner on integrated water management in the world due to the introduction of a number of generally agreed principles and concepts into a binding regulatory instrument (Commission 2007). Consequently, the WFD and European IRBM in general have become a source of inspiration for water management reforms elsewhere, most notably in China. The Chinese water authorities have shown interests in adopting the WFD to fight the river pollution problems and manage flood risks that they are experiencing, e.g., in the cases of the Yellow River and the Yangtze River.

## The Negotiated Approach in Bangladesh

A third approach is emerging in the form of increasing stakeholder dialogue, both in developing countries and in developed countries. In the context of developing countries, it got a name "the negotiated approach" (Both Ends and Gomukh 2005). We use the example of Bangladesh. In 1990, the Flood Action Plan, a nationwide water development program, was launched in Bangladesh. Drawn up in reaction to disastrous floods in the late 1980s, it involved huge foreign support, both financially and technologically. The program was coordinated by the World Bank, at the request of the government of Bangladesh. Consequently, infrastructural works like dikes, polders, and sluices were planned and implemented in the early 1990s. In the beginning the effects seemed positive in that the measures made possible the introduction of

high-yielding rice varieties. Recently, however, more adverse impacts have become visible. Increasing siltation in canals and rivers led to the lowering of water tables in (wet)lands, heightening of river beds, and reduced conveyance capacity of rivers, resulting in long-term waterlogging, especially in the southwest. Consequently, farmers could not use their rice fields, while roads and villages remained under water. Severe siltation has caused whole rivers to disappear, though the Bangladesh Water Development Board (BWDB) is planning to restore these. However, measures are implemented only slowly, while even local communities feel that constant and repeated dredging of rivers and canals is not a sustainable solution.

By sheer frustration and led by a local NGO, a group of farmers in the southwest of Bangladesh started to implement Tidal River Management (TRM), based on opening polders periodically to tidal flows. People see to it that the river drops its silt in a polder, which is temporarily put under water, instead of in its bed. Since this was done without permission of the central authorities, the communities and NGOs involved are now in conflict with the Bangladeshi government, represented by the BWDB. One of the issues is that the BWDB wants to remain in control of water management, i.e., the opening and closing of sluices and the decisions about how much water is flowing where. The local farmers and NGOs, however, do not trust the central government on that. They argue that government officials in Dhaka do not care about the problems of the farmers and that decisions on water control should be taken locally anyhow. International donors are mediating, trying to reach agreement and solutions by means of negotiations, but local communities, remembering their alignment with the BWDB in earlier projects, are suspicious and not much progress has been made with this negotiated approach.

This approach in which different stakeholders are brought together in order to create mutual understanding and find common ground for solutions meeting their diverging interests and values is increasingly used to overcome situations in which different stakeholders are at loggerheads, like in the "Green Water Credits" approach, which supports upstream land users to invest in effective soil and water conservation practices, optimizing both upstream and downstream water use (ISRIC 2013).

## Design for Values in Water Engineering and Management

A variety of interests and values is connected with water:

1. Functional interests and objectives connected to the multipurpose character of water systems
2. Trans-boundary interests and values related to territorial divisions, like nations and like group identities and cultural areas
3. Social objectives and values determining the conditions under which water systems have to operate, such as safety, sustainability, and justice
4. Sociopolitical values and characteristics, generally involving cultural uniformity versus diversity, centralization in management versus decentralization, and private versus public ownership

How to reconcile all these conflicting values and interests in water resources development and management? Several methods constituting a basis for comparison, consideration, and decision-making are available, including:

(a) Cost-benefit analysis
(b) Multi-criteria analysis
(c) Defining boundary conditions or thresholds
(d) Considering technological alternatives or innovations (Van de Poel and Royakkers 2010)

For decision-making in the framework of water engineering and management, these methods are all more or less in use. The technical-economic approach relies on multipurpose works, in which all innovative capacity is directed towards developing existing technologies further and further, culminating in bigger and bigger dams and reservoirs, e.g., the Three Gorges Dam in China and the Itaipu Dam in Brazil and Paraguay. These projects are expensive and financed by their return, at least in part. Consequently, the water is predominantly distributed on economic principles, where most profit could be secured (cost-benefit analysis). This approach is therefore strongly connected to utilitarianism, as it is optimizing utility, taking it as a self-evident value defined from one central perspective. In this approach, in which one dominant actor defines what utility consists of, there is little room for including the perspectives of different stakeholders. Since these perspectives imply the maximization of different values including them would turn the concept of utility into a "container concept". In case of a reservoir, for instance, for the fishers utility could mean the maximization of food security for their families. For the government it could mean the maximization of energy production. For those using the reservoir for touristic purposes it could mean the maximization of rest or fun. Each time the concept utility contains a different content in terms of values served.

The integrated management approach usually meets the values and interests of all stakeholders through negotiations and trade-offs (e.g., multi-criteria analysis), though economic considerations play a role as well. Both approaches have a strong top-down orientation, while the influence of engineers is strong, be it civil engineers or policy engineers. The negotiated approach is based on negotiations and trade-offs as well; though from a bottom-up perspective, the needs and perspectives of local people are leading (constitute a boundary condition). The more stakeholders are involved, the more value perspectives need to be taken into consideration. For an integrated management approach, the evaluation of such different value perspectives is an attractive instrument, as part of a multi-criteria analysis done by experts. In the negotiated approach it is the participants in the dialogue themselves who need to conduct such analyses, in an endeavor to find some form of common ranking. The arguments for one particular solution may differ, while this particular solution is still attractive from different perspectives.

All approaches take technological alternatives and innovations into account, be it other megaprojects, aimed at connecting and redirecting rivers, e.g., in Spain the

National Water Plan and in China the South-North Water Transfer Plan (technical-economic approach), increasing the storage capacity of rivers (IWRM), or endogenous technological and management solutions (negotiated approach). However, in view of addressing present-day water problems and reconciling all conflicting values and interests involved, the technical-economic approach is facing severe limitations, especially when it comes to big dams, aligned as it is to a linear centralized approach. The other two approaches are much more promising, though we think that all approaches would gain from a systematic and explicit design for values perspective. In the next paragraphs we will indicate how.

Design for values has been developed from a computer design perspective (see chapter "► Design for Values in ICT"). Nevertheless, the problem is analogous. As computer programs became more complicated and were implemented in more complex organizations, it appeared that different stakeholders would pose different value priorities and that trade-offs and agreements needed to be found between different stakeholders defending different value priorities. Without such embedding within organizations, beautifully designed network programs were not accepted and did not function well. Triggered by this development, ethicists as well began to think of moral deliberation as an optimization process in analogy to a design process (Whitbeck 1998). The value-sensitive design approach towards the design of computer software architecture is in itself an endeavor to take such different value perspectives into account (Van de Poel and Royakkers 2010).

Friedman et al. (2006) try to cope with this complexity by a three-step analysis (conceptual, empirical, design, see chapter "► Value Sensitive Design: Applications, Adaptations, and Critiques"). In comparison to the traditional methods used in ethical analysis and evaluation, the new method brings two important innovations, one implicitly and one explicitly. Implicitly, it allows for a more historical and contingent interpretation of particular values. These are not static like from all eternity, but the meaning of loyalty, trust, well-being, and many of such moral notions may need different interpretations and specifications from context to context. There is a felt need to do field research to find out this contextual meaning of such notions. Explicitly, the design for values method installs a deliberate trade-off like usually is the case in any design problem between different demands and value preferences. The more complex the problem may become and the more participants need to be drawn in to find a solution, the more differentiated value preferences need to be accommodated. But this does not only make the entire process more complex, it also leads to innovative ideas that in unexpected ways may meet the design requirements. In that sense it leads to technical innovation as well, by discovering and creating options which otherwise would not have been envisaged (Van de Poel and Royakkers 2010). Finding trade-offs, agreements, and compromises on water issues could follow the same path.

Constructing and implementing such a design for values perspective might meet the current search for solving conflicts and competition in water use, which will only increase in number and intensity, at a more fundamental level. This first requires an analysis of the concepts by which the debate is framed from different sides. What initially merely seem to be differences of interests, for instance,

between farmers, fishermen, and cattle herders or between big cities and rural areas, at a closer look will also appear to include different value priorities: water security for large city populations over against a basic needs subsistence economy for rural farmers, shrimp production for export purposes creating revenues for the central government (important for maintaining peace between different political factions) versus rice and fish production for the family (important for the good health of children), etc. The last example is from Bangladesh. Such an analysis leads to an understanding and tackling of water problems, ultimately, in terms of multiple values, which necessitate negotiation and dialogue about value priorities, like in the negotiated approach (cf. Glenna 2010). Depending on the trade-offs found and the level of agreement reached, technological options can be designed and adapted to this shared support base. Ethical analysis and social debate and dialogue aim at designing water works and management in new ways, including institutional (re) design and social experimentation. Stakeholder involvement is a precondition and it suggests efforts at various levels, both top-down and bottom-up. Central planning will always occupy an important place, especially in countries like China, but for the sake of using the full potential of all technological and management options at our disposal, the need makes itself felt to include more and more decentralized options and, consequently, involve local participants in the design of future water provisions. Increasing complexity of water problems and solutions and the necessity of stakeholder participation make it impossible to find the one best solution from the perspective of one central point of control, from a supposedly self-evident but actually naive utilitarian perspective. This is the more so if it is taken into account that a large part of the solution of water problems consists in a long chain of many small-scale and local solutions: a small dam in a particular village, a small diversion of a little river to fill a tube well, etc.

We could, in turn, enrich the design for values perspective by using some insights and methods which have been developed in the domain of technology assessment, which maps and evaluates (negative) impacts of new (emerging) technologies (see chapter "► Technology Assessment and Design for Values"). Technology assessment has a strong ethical dimension, though its approach is mainly sociological, relying heavily on stakeholder engagement. A simple approach for TA embraces the following steps: (1) defining problem and research questions, e.g., for whom, problem versus technology oriented; (2) exploring and fore sighting technological developments; (3) technology impact assessment; (4) normative judgment; and (5) generating improvement options (Smit and Van Oost 1999). Especially, constructive technology assessment (Schot and Rip 1997; Quist 2007) may be meaningful in the further development of design for values. This method, based on constructivist theories of technology development (Bijker et al. 1987), aims at the participation of the stakeholders right from the start, enabling them to co-shape new technologies.

Design for values emphasizes the exploration of the different values of the stakeholders and their feedback on the design of the technology, but following constructive technology assessment, it could move beyond stakeholder consultation and more explicitly involve stakeholder participation in the construction of not only the technical but also the moral solutions required: value trade-offs and

specifications. The important point to be made here is that the interaction between the different stakeholders in the debate and dialogue process has a surplus value beyond their initial partial and biased preferences. That means that in an open discussion, creative technological and management solutions can be found which would otherwise not have been discovered. This does count not only for management and technology but also for the context-dependent concretization of the values involved. Open moral deliberation has an impact on the understanding and concretization of these value differences (Kroesen and van der Zwaag 2010). Even competing moral values may be reconciled not only by a compromise but by alternating between them in the right order between different parties (consider the use of water and land by fishermen, farmers, and herders). The process of dialogue and discussion may be difficult and time-consuming, but the results may turn out to be more sustainable and enduring than the implementation of quick linear central solutions. Below, we will explore and show the surplus value of integrating technology assessment views in the design for values perspective for the three approaches to water problems we started out to discuss.

## Revaluing the Technical-Economic Approach

Building big dams and other structural works could greatly profit from the step-by-step plan outlined above. Suggestions go in this direction, without, however, explicitly making an inventory and analysis of value backgrounds (Reddy 1999). In the case of India, the earlier cited report suggests that local sustainable water sources should be tapped and maintained such as rain water harvesting and the use of check dams and many more options. This advice points into the direction of decentralization of water provisions, intersectoral cooperation, and institutional reorganization. The report concludes: "If these options are properly taken into account, there is little justification of large dams as option for urban water supply" (Assessment 1999, p. 2). Such a policy would require more management and more negotiation and as a consequence requires to take into account a host of different values and value priorities from those different actors in order to achieve a support base for common action. This trend towards including multiple perspectives and different stakeholders puts (a diversity of) values and consequently design for values at the center of policy and technology debates.

A historical example in which this procedure has been followed, though not as a procedure but as a developing practice, is the construction of the Eastern Scheldt storm surge barrier between the islands Schouwen-Duiveland and Noord-Beveland in the Dutch province of Sealand (1976–1986). It was the showpiece of the ambitious Delta works series of dams, designed in response to the North Sea Flood of 1953. The 9-km-long barrier was initially designed as a closed dam, but after public protest by the local oyster and mussel farmers for economic reasons and, later, environmental groups from ecological values, huge sluice-gate-type doors were installed over 4 km. These doors can be closed if weather conditions are threatening, thus protecting the land from flooding. However, because usually

they are open, the saltwater marine life behind the dam remained preserved and oyster and mussel farming could be continued (Lintsen 1998; Disco and van der Vleuten 2002). In this case deliberate value trade-offs in the design of the dam were made not just from a multi-criteria perspective but more as a result of pressure from below and negotiations between different stakeholders, including biologists. The solution was creative: a technical option which realized different contradictory values and interests. The result was a high-tech engineering work, which despite being expensive turned out to be an asset, because it largely contributed to the international reputation of the companies involved.

## Revaluing IWRM

IWRM, e.g., in the form of the European Water Framework Directive, has shown positive results already, though it also could lead to tensions with existing water traditions (Ravesteijn and Kroesen 2007). The WFD has a strong "negotiating content" and it is suffused with a spirit of "deliberation, education and collaboration" (Bohensky et al. 2009). In its further implementation it could easily include values and value trade-offs in river basin management and form the basis for further reflection on a more deliberate introduction of values into the debate. Challenge would be to avoid and fix tensions that could easily rise between top-down directives and bottom-up wishes.

Transferring the WFD to other countries and parts of the world introduces new challenges, as it very clearly presumes a specific institutional context (de Jong et al. 2002). Current efforts to model Chinese river basin management after the WFD, especially in the basin of the Yellow River, constitute an example (Song et al. 2009). River management in China differs a lot from Europe (see Table 2).

A complete transfer of the WFD to China seems impossible. Three points are relevant here (Song et al. 2009):

1. A striking difference is the political structure as well as political tendencies, which determine developments and possibilities in the water domain. China and Europe are representing two archetypical organizational models for implementing IRBM, i.e., the authority model and the coordination and negotiation model. Interestingly, however, a tendency to convergence can be noted, though great differences still remain.
2. It makes sense to manage rivers in regard to hydrological boundaries. The problem, however, is how to distribute power between China's multiple administrative water management agencies, including river basin commissions.
3. The involvement of local stakeholders in basin-level planning and actions was right from the start a main point in Europe, naturally related to the political organization of decision-making and policy implementation in Europe. How to realize such in present-day China? In general, China's progress in realizing IRBM depends on the public awareness of environmental problems in relation to economic growth and the development of a civil society (Song et al. 2009, 2011).

**Table 2** Comparison of contemporary water management between Europe and China

| Aspect | WFD | Chinese river management |
|---|---|---|
| Objectives | Good condition of all waters (including surface and groundwater) | Water conservation and pollution prevention (mainly surface water) |
| Scope of planning | River basin planning, update river basin plan every 6 years | Combination of river basin planning and regional administrative planning |
| Pollutants management | Combined approach towards control of both point and nonpoint source pollutants | Focus on the control of point source pollutants; no effective measures for nonpoint source pollutants |
| Decision-making | Top-down and bottom-up, centralized and decentralized management | Top-down, centralized management with strong and varied hierarchy |
| Role of RBCs | Responsible for all water-related issues at river basin level | Mainly responsible for water quantity management, e.g., flood control and water allocation |
| Water allocation and water rights | Controls on water abstraction and groundwater recharge; member states' own policies specify water rights | Rational allocation to alleviate the upstream-downstream conflicts; ambiguous water rights at regional and local levels |
| Water pricing | Full cost recovery and cost-effective provisions to be taken into account | Preliminary research on water price, its components, and measurement |
| Public participation | Getting citizens more closely involved in river basin management activities from early stages | Insufficient stakeholder participation in water planning and management as well as flood control |

There are many successful examples of IRBM or IWRM in general, in the Netherlands and elsewhere (Dubbelman 1999; He 2011). The challenge, however, is to analyze conflicting interests at a deeper value level, in order to easier find common ground and social support. A promising methodology in this regard is Q methodology, which is aimed at mapping underlying "stakeholder perspectives" and thus searches for common ground at fundamental levels (Cuppen 2010). Although stakeholder participation in the moral and technical construction of the solution is difficult in China, it is clear that in this case too it might lead to more differentiated and complex solutions, but also and for exactly that reason to more effective and acceptable solutions.

## Revaluing the Negotiated Approach

The Negotiated Approach as developed and applied in Bangladesh is promising, but unlike some other cases (Both Ends and Gomukh 2005), it is constrained by the under-institutionalized character of the surrounding sociopolitical context. It should be embedded in a broader sociopolitical context which requires institutional design and development, as well as experimentation (Ravesteijn et al. 2011). The strongly collectivist culture of the Bangladeshi system of governance appears to be an obstacle in this regard. Although there are elections, the respective political parties

serve their own electorates. There is no culture of negotiation and compromise; time and again it is "we" against "them." This is also manifest at the local level and between the different departments. The technology-minded Bangladesh Water Development Board (BWDB) does not consider uneducated farmers as negotiation partners, while farmers have no confidence whatsoever in the officials. The different parties, therefore, remain at loggerheads.

The lack of trust between the different parties needs to be overcome. A change towards a culture of give and take between parties seems a necessary condition. Images of friend and foe now dominate all actions and expressions. A process of meaningful dialogue, in which people can reach compromises, could redress this. Building trust between government and farmers is essential. Once the dialogue will be opened, immediately a diversity of values and value priorities will urge itself on the debate. The BWDB is primarily technology oriented and in support of central control: it does not take the farmers seriously as partners, is focused on technical solutions, and does not want to share power and control. The farmers are concerned about their land and the well-being of their families and want to be considered as participants in finding solutions, but on the other hand for the moment, they are not yet prepared to open up towards government officials in order to find negotiated trade-offs. Environmental concerns, bureaucratic concerns, and economic concerns compete with each other. Values like openness or distrust, egalitarianism or authority, local cooperation or central control, and confidence in management or in technology may not be explicitly articulated, but exactly for that reason their impact on the process is large. For the moment the farmers resort to collective action and remain at loggerheads with the equally intransigent water authorities, whereas recent research has shown that in actual fact the opinions about practical solutions and trade-offs are not so remote as the heat of the debate may suggest (Brockhus and Das 2010). In this case it is quite clear that the solution is not only dependent on value trade-offs and thus cannot be reached by a multi-criteria analysis or by a desk study, but that stakeholder participation is necessary in constructing the proper moral and technical solution.

Consequently, bottom-up water control, like the negotiated approach in Bangladesh, should be embedded in a national framework. The Dutch water history provides a good example where this took place, paving the way for a successful water engineering and management approach. In the case of the "big rivers problem" in the Netherlands, the traditional water associations were unable to come up with a solution; that required too much from their abilities to cooperate. After several vain attempts made by these associations, the problem of flooding from the Rivers Rhine and Meuse could only be solved by the national water agency (Rijkswaterstaat), which was founded in 1798 after French example (Lintsen 1998; Kaijser 2002). Ever since, the combined bottom-up and top-down approach in Dutch water interference has kept the country safe seeing to it that its inhabitant could keep their feet dry, more or less at least, and with new challenges and new water approaches (like building with the water) having emerged as a result of climate change and other developments.

## Conclusions and Future Work

There has been a notable shift from framing water problems as merely technical problems towards involving management issues and social interests. Currently, a new shift is underway, aimed at involving different values in water engineering and development. This is not a complete surprise, since the complexity of the problems and the limitations of often applied large-scale centralized solutions urge the inclusion of a great and increasing number and variety of stakeholders. These stakeholders happen to frame their perspectives by different interests and diverse values. We gave a number of examples. This development requires new methods of design and development, in which negotiation and dialogue about value priorities are extraordinarily relevant, as well as institutional (re)design and social experimentation. As the tendency of the debate around water issues is already moving towards reconciling different value priorities (or failing to do so), the quality of the debate will only gain by consciously and deliberately putting different values and stakeholder perspectives on the agenda. This could improve existing water engineering and management approaches, reconciling top-down and bottom-up viewpoints. Stakeholder participation is of utmost importance. Negotiated approaches, like in Bangladesh, are very promising, though the traditional engineering and management approaches remain relevant.

This paper has discussed three approaches of water engineering and management, assessing their performance and potential in relation to a design for values perspective. Obviously, the third approach we have distinguished comes close to this new way of dealing with water problems, as it takes the water users and stakeholders as points of departure. However, even in this case, value considerations could be included more explicitly, ultimately giving rise to an elaborate design for values approach. How would that look like? We have argued that design for values could benefit from technology assessment concepts and tools, especially when it comes to integrating stakeholder participation in the ethical analysis and even more in the construction of moral and technical solutions. A specific and systematic way to deal with water issues could include the following method or step-by-step plan:

1. Find out what problem has to be solved and which goals have to be served. This could be done in consultation with parties that have put the problem on the agenda.
2. Make an elaborate stakeholder analysis, mapping all relevant parties beside the initiating actors, their perceptions of the problems and solutions, their interests and values, their arguments and their lines of reasoning, as well as their resources.
3. Make an ethical-philosophical conceptual analysis of the diversity of perceptions, interests, and values as well as the arguments and lines of reasoning, aimed at deriving and constructing fundamental, underlying values and perspectives.
4. Based on the outcomes of step 3, make a list of all alternative solutions, considering each and every function that needs to be dealt with. In case of a

big dam, e.g., these could also include combinations or programs of small alternative works. This is the design phase in the process, in which scenarios are developed with different trade-offs and equilibria of values and interests.
5. By means of stakeholder consultation and deliberation, based on this package of moral and technical alternatives, make an integrated impact assessment of each alternative (ecological, social, economic, strategic, etc.), considering basic perceptions, values, and perspectives of the stakeholders.
6. Make an additional (impact) analysis of the implementation strategies that could or have to be used, again considering basic perceptions, values, and perspectives of the stakeholders, which result from stakeholder consultation and deliberation.
7. Select a solution in consultation and cooperation with the stakeholders. Urgent problems, different value traditions, and physical and institutional constraints should find an optimal trade-off in open deliberation, as much as is possible within the existing political framework.
8. Finally, implement the selected solution in consultation and cooperation with the stakeholders. Indicate who should do what.

As particularly comes to the fore in the Bangladeshi case, the explicit and open discussion of different values, instead of positioning them as mutually exclusive, opens a new field for experimentation and research. It shows the necessity of debate and dialogue in order to find more creative and differentiated solutions and gain a stronger support base for common action. It needs to be emphasized that the way forward in this type of problems cannot be found by desk research and theory development only. Although theory can summarize and systematize past experiences and thereby prevent making the same mistakes once more, it cannot show in advance what only the experiment of life itself can teach us. For future research social experimentation is a requirement, but it should not be blindfold. It should be accompanied by thorough reflection and be conducted as a sort of action research, involving and bringing together many stakeholders.

## Cross-References


- ▶ Conflicting Values in Design for Values
- ▶ Design for the Value of Sustainability
- ▶ Design for Values in ICT
- ▶ Technology Assessment and Design for Values
- ▶ Value Sensitive Design: Applications, Adaptations, and Critiques


## References


Assessment of Water Supply Options (1999) South Asian network on Dams, Rivers and People. Nov 1999

Bijker WE, Hughes TP, Pinch TJ (eds) (1987) The social construction of technological systems: new directions in the sociology and history of technology. MIT Press, Cambridge, MA

Bohensky E, Connell D, Taylor B (2009) Experiences with integrated river basin management, international and Murray Darling Basin: lessons for northern Australia. Northern Australia land and water science review full report. http://www.nalwt.gov.au/files/Chapter_22-International_experience-lessons_for_northern_Australia.pdf

Both Ends and Gomukh (2005) River basin management: a negotiated approach. Both Ends, Amsterdam, and Gomukh, Pune

Bressers H, Kuks S (eds) (2004) Integrated governance and water basin management. Kluwer, Dordrecht/Boston/London

Brockhus P, Das P (2010) The use of Q. Methodology: a social approach regarding the waterlogging problem in the south coastal region of Bangladesh. Students internship report TU Delft

Cech TV (2009) Principles of water resources: history, development, management, and policy. Wiley, New York

Civil Engineering (2013) http://www.britannica.com/EBchecked/topic/119227/civil-engineering. Oct 2013

Commission of the European Communities (CEC) (2007) Towards sustainable water management in the EU; First stage in the implementation of the water framework directive 2000/60/EC. Brussels

Cuppen EHWJ (2010) Putting perspectives into participation: constructive conflict methodology for problem structuring in stakeholder dialogues. Vrije Universiteit, Amsterdam

Dam (2013) Wikipedia. Oct 2013

de Jong M, Lalenis K, Mamadouh V (eds) (2002) The theory and practice of institutional transplantation; experiences with the transfer of policy institutions. Kluwer, Dordrecht

Disco C (2002) Remaking "nature": the ecological turn in Dutch water management. Sci Technol Human Values 27(2):206–235

Disco C, van der Vleuten E (2002) The politics of wet system building: balancing interests in Dutch water management from the middle ages to the present. Knowl Technol Policy 14(4):21–40

Dubbelman H (1999) Maatschappelijke golven in de waterbouwkunde. Delft University Press, Delft

Europe Environmental Agency (EEA) (2009) River basin management plans and programme of measures. http://www.eea.europa.eu/themes/water/water-management/themes/water-management/river-basin-management-plans-and-programme-of-measures. Accessed 11 Apr 2009

European Commission (DG Environment) (2008) Water note no. 2, Wise 2008. ec.europa.eu/environment/water/water-framework/pdf/water_note2_cleaning_up.pdf

European Commission (EC) (2000) Introduction to the new EU water framework directive. http://ec.europa.eu/environment/water/water-framework/overview.html

European Water Conference (EWC) (2009) http://ec.europa.eu/environment/water/index_en.htm

Floods Directive (2007) http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:288:0027:0034:EN:PDF

Friedman B, Kahn PH Jr, Borning A (2006) Value sensitive design and information systems. In: Zhang P, Galletta D (eds) Human-computer interaction in management information systems: foundations. M.E. Sharpe, Armonk/London, pp 348–372

Glenna LL (2010) Value-laden technocratic management and environmental conflicts: the case of the New York City Watershed Controversy. Sci Technol Human Values 35:81–112

He J (2011) A policy analysis approach to making strategic decisions for sustainable development of Harbin Eco-corridor system. Master thesis Harbin Institute of Technology and TU Delft

Hermans L (2005) Actor analysis for water resources management. Eburon, Delft

Hoekstra AY, Huynen M (2002) Balancing the world water demand and supply. In: Martens P, Rotmans J (eds) Transitions in a globalising world. Swets & Zeitlinger, Lisse, pp 17–35

International Rivers (2013) http://www.internationalrivers.org/campaigns/reservoir-emissions. Oct 2013

ISRIC (2013) http://www.isric.org/projects/green-water-credits-gwc. Oct 2013

Kaijser A (2002) System building from below: institutional change in Dutch water control systems. Technol Cult 43(3):521–548
Kates RW, Burton I (eds) (1986) Themes from the work of Gilbert F. White. Geography, resources and environment, vol 2. The University of Chicago Press, Chicago
Kissling-Näf I, Kuks S (2004) The evolution of national water regimes in Europe. Kluwer, Dordrecht/Boston/London
Kroesen JO, van der Zwaag S (2010) Teaching ethics to engineering students: from clean concepts to dirty tricks. The impact of practical circumstances and personal relationships on ethical decision-making. In: Goldberg DE, van de Poel IR (eds) Philosophy and engineering: an emerging agenda. Philosophy of engineering and technology, vol 2. Springer, Dordrecht/New York/Berlin, pp 127–137
Lintsen H (1998) Twee Eeuwen Rijkswaterstaat, 1798–1998. Walburg Pers, Zaltbommel
McCully P (2003) Big Dams, Big Trouble. http://www.highbeam.com/doc/1G1-99232389.html
McCully P (2011) Dam decommissioning. http://www.internationalrivers.org/node/571
Quist J (2007) Backcasting for a sustainable future: the impact after 10 years. Eburon, Delft
Ravesteijn W (2002) Een ingenieur met visie. Prof. Dr. Ir. Willem Johan van Blommestein (1905–1985). Tijdschrift voor Waterstaatsgeschiedenis 11(1):6–11
Ravesteijn W, Kroesen O (2007) Tensions in water management: Dutch tradition and European policy. Water Sci Technol 4:105–111
Ravesteijn W, Hermans L, van der Vleuten E (2002) Water systems. Participation and globalisation in water system building. Knowl Technol Policy 14(4):1–163
Ravesteijn W, Kroesen O, Firoozyar F, Song X (2011) River systems in transition: pathways and stakeholder involvement. WIT transactions on ecology and the environment 126:327–339. Also published in: Brebbia CA (ed) River basin management VI. WIT Press, Southampton; and in Transactions of the Wessex Institute, eLibrary. http://library.witpress.com/pages/PaperInfo.asp?PaperID=20543
Reddy AKN (1999) Big dams: a fresh approach. http://www.narmada.org/archive/hindu/files/hindu.19990920.05202524.htm
Resources Directive (2006) http://www.central2013.eu/fileadmin/user_upload/Downloads/Document_Centre/OP_Resources/08_Directive_2006_118_EC.pdf
Ricketts JT, Loftin MK, Merritt FS (2004) Standard handbook for civil engineers. McGraw-Hill, New York/Chicago/San Francisco/Lisbon/London/Madrid/Mexico City/Milan/New Delhi/San Juan/Seoul/Singapore/Sydney/Toronto
Sachs J (2008) Common wealth: economics for a crowded planet. Penguin, New York
Schot J, Rip A (1997) The past and future of constructive technology assessment. Technol Forecast Soc Change 54(2):251–268
Smit WA, van Oost ECJ (1999) De wederzijdse beïnvloeding van technologie en maatschappij. Een technology assessment-benadering. Coutinho, Bussum
Song X, Ravesteijn W, Wennersten R (2009) The 2000 EU water framework directive and Chinese water management: experiences and perspectives. WIT transactions on ecology and the environment 124:37–46. Also published in: Brebbia CA (ed) River basin management V. WIT Press, Southampton; and in Transactions of the Wessex Institute, eLibrary. http://library.witpress.com/pages/PaperInfo.asp?PaperID=20543
Song X, Ravesteijn W, Mulder K, Frostell B, Wennersten R (2011) Transition in public participation in Chinese water management. Eng Sustain ES1:71–83
van de Poel I, Royakkers L (2010) Ethics, technology and engineering: an introduction. Wiley-Blackwell, Malden Oxford, Chichester
van de Ven GP (2004) Man-made lowlands: history of water management and land reclamation in the Netherlands. Matrijs, Utecht
Whitbeck C (1998) Ethics in engineering practice and research. Cambridge University Press, Cambridge

# Index





J. van den Hoven et al. (eds.), *Handbook of Ethics, Values, and Technological Design*,
DOI 10.1007/978-94-007-6970-0