Philosophy of Engineering and Technology

Peter Kroes
Peter-Paul Verbeek *Editors*

# The Moral Status of Technical Artefacts

Springer

# The Moral Status of Technical Artefacts

# Philosophy of Engineering and Technology

VOLUME 17




For further volumes:
http://www.springer.com/series/8657

Peter Kroes • Peter-Paul Verbeek
Editors

# The Moral Status of Technical Artefacts

Springer

*Editors*
Peter Kroes
Department of Philosophy
Delft University of Technology
Delft, The Netherlands

Peter-Paul Verbeek
Department of Philosophy
University of Twente
Enschede, The Netherlands

ISSN 1879-7202 ISSN 1879-7210 (electronic)
ISBN 978-94-007-7913-6 ISBN 978-94-007-7914-3 (eBook)
DOI 10.1007/978-94-007-7914-3
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013958227



Printed on acid-free paper

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

# Contents

# Chapter 1
# Introduction: The Moral Status of Technical Artefacts

**Peter Kroes and Peter-Paul Verbeek**

In recent decades, discussions about the question of how to morally assess technology and its influence on human beings have taken a new, intriguing twist. Issues about the moral status of technology—in the sense of whether technology itself or its influence on human life may be evaluated as morally good or bad—have a long history. But recently, various proposals have been put forward to ascribe some form of moral agency to technology, more in particular to technical artefacts.

A common idea underlying these proposals is that the way technical artefacts influence human behaviour cannot be captured and understood by taking technical artefacts to be merely passive instruments to be used at will for morally good or bad purposes. Instead, technical artefacts are supposed to play a much more active role with regard to humans and to actively shape the human condition. For a better understanding of this active role of technology in human life, technical artefacts in their use by and in their associations with human beings are regarded as being agents of some sort. Being 'agents' rather than simply passive instruments, technical artefacts may actively influence their users, changing the way they perceive the world, the way they act in the world and the way they interact with each other. In the most far-reaching of these proposals, technical artefacts as agents are taken to be susceptible to moral assessment: more or less similar to human beings, technical artefacts themselves and their actions may be qualified as morally good or bad. In this view, the image of technical artefacts being morally neutral passive instruments is replaced with an image in which technical artefacts figure as some sort of moral agents.

P. Kroes (✉)
Department of Philosophy, Delft University of Technology, Delft, The Netherlands
e-mail: p.a.kroes@tudelft.nl

P.-P. Verbeek
Department of Philosophy, University of Twente, Enschede, The Netherlands
e-mail: p.p.c.c.verbeek@utwente.nl

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_1,

These proposals to attribute some form of moral agency to technical artefacts may be taken to be the most recent offspring of an on-going attempt to interpret the role of technology in relation to what is often referred to as 'the good life'. What is the moral status or the moral significance of technical artefacts? Discussions about the moral agency of technical artefacts have to be put against a background of diverging views on how to evaluate the impact of technology on human life. The roots of these diverging views can be traced back far in to the past to stories that express deeply seated positive and negative sentiments towards technology. Whereas in the Greek story of Prometheus, technology in the form of fire is portrayed in a positive way (a precious gift for humankind, stolen by Prometheus from the Gods), the story of Icarus flying to the sun is a story of hubris, where technology leads humans astray by seducing them to overconfident and reckless undertakings. The wings he made out of feathers and wax required Icarus to find the right middle; if he flew too low, the feathers would absorb water, and if he flew too high, the wax would melt. But he flew too high, and crashed into the sea. Bacon's story of the house of Salomon extolls the positive impact of science and technology on the life of the inhabitants of the isle of Bensalem, whereas in Butler's story the inhabitants of Erewhon have outlawed machines because of their negative effects.

The following story, quoted by Heisenberg from the writings of the Chinese sage Chang Tsi living about two and a half thousand years ago, exemplifies an outspoken negative attitude towards technology (Heisenberg 1958, pp. 102–103):

> When Tsi Gung came into the region north of the river Han, he saw an old man busy in his vegetable garden. He had dug ditches for watering. He himself climbed into the well, brought up a container full of water in his arms, and emptied it. He exerted himself to the utmost, but achieved very little. Tsi Gung spoke: "There is an arrangement with which it is possible to fill a hundred ditches with water every day. With little effort much is accomplished. Wouldn't you like to use it?" The gardener rose up, looked at him and said, "What would that be?" Tsi Gung said, "A lever is used, weighted at one end and light at the other. In this way water can be drawn, so that it gushes out. It is known as a draw-well." At that, anger rose up in the face of the old man and he laughed, saying, " I have heard my teacher say: 'When a man uses a machine he carries on all his business in a machine-like manner. Whoever does his business in the manner of a machine develops a machine heart. Whoever has a machine heart in his breast loses his simplicity. Whoever loses his simplicity becomes uncertain in the impulses of his spirit. Uncertainty in the impulses of the spirit is something that is incompatible with truth.' Not that I am unfamiliar with such devices; I am ashamed to use them."

What is particularly interesting about this story is that the negative attitude of the old man towards technology is not based on his negative moral assessment of particular goals that may be achieved with the help of technology, or on risks associated with technology. It is not that he thinks that people will use technology in morally bad ways and that that is the reason he is against the use of technology. Something much more fundamental is at stake. He is ashamed to use technology because of a particular feature of the use of technology in general, namely that this use corrupts the spirit of its user. That is the reason why in his opinion the use of technology does not contribute to or, more strongly, is not compatible with leading the good life. It is an inherent feature of technology (its machine-like character) that makes the use of technology morally bad.

It is not difficult to discern in these stories the seeds of ideas that are at issue in recent debates about the moral status of technical artefacts, in particular about their moral agency: Is the morally positive or negative impact of technology due to the way humans use technology or to the way technology (actively) conditions human life? Are humans to be praised or blamed for the impact of technology on their efforts to bring about the good life, is it technology itself, or is it the interaction between human users and technological artefacts? Is technology itself a curse or a blessing when it comes to living a good life? It is rather curious (and telling!) that in stories that stress the positive role of technology there appears to be a tendency to praise humans for their wise use of technology, whereas in stories that stress its negative role technology often takes the blame by depicting it in some form of a bad and uncontrollable demon as in the story about the Golem. Although the Golem (technology) is a creation of humans, it starts to lead a life of its own, that is, to act as an autonomous agent and it is that autonomous agent that is to blame for its negative moral impact on human life.

Of course, it is one thing to depict technology metaphorically as a (morally good or bad) agent, as is done in some of the stories mentioned above, it is quite another thing to attribute moral agency to technology and technical artefacts in a literal sense. But exactly that is at stake in the recent debate about the moral agency of technical artefacts. One of the most interesting aspects of the current discussion about moral agency and technical artefacts is it has shifted from a debate about whether the impact of technology on leading the good life is to be evaluated morally in a positive or negative way, to a debate about an underlying, more fundamental issue, namely the issue whether it makes sense to qualify technical artefacts in some literal sense as morally good or bad in themselves, more in particular as morally good or bad agents.

According to traditional ethical thinking only acts of agents or agents themselves may qualify as entities that can be assessed morally, not the things made or used by agents. It does make sense to morally evaluate the act of making or using a particular technical artefact or natural object, but the technical artefacts and natural objects themselves fall outside the domain of what is susceptible to moral judgment. From this point of view, it would be tantamount to committing a category mistake to call a speed bump morally good or a gun morally bad. The same applies to the characterization of technical artefacts as morally neutral objects by defenders of the moral neutrality thesis. According to traditional ethical theories such a characterization is confused because technical artefacts do not admit of any moral qualification by themselves, whether good, bad or neutral. The fact that technical artefacts may be used for morally good or bad purposes does not make them morally neutral objects. Anybody who thinks that technical artefacts are morally good, bad or neutral erroneously takes technical artefacts as objects of moral evaluation instead of acts with or related to these artefacts.

From a traditional ethical point of view, therefore, the most obvious way to try to draw technical artefacts into the domain of the moral is by assuming that they act or are agents in ways that in a morally relevant manner are similar to human acts or human agents. This approach leads to the question whether the observation

that technical artefacts actively influence human behaviour and human beings in far-reaching ways may be interpreted as implying that technical artefacts may be attributed a morally significant form of agency. This may be the most obvious way, but it surely is not an easy way. Claims about moral agency of technological artefacts have revolutionary pretensions. After all, after the Enlightenment shifted the source of morality from God to humans, these claims want to move it one step further: from humans to material things. This 'material turn' in ethics raises many questions, though. Is the conclusion that material things influence human actions reason enough to actually attribute morality to materiality? Can material things be considered moral agents, and if so, to what extent? And to what extent can artificial moral agents be constructed with the help of information technology? The attribution of some form of moral agency to technical artefacts not only requires a rethinking of the notion of agency but also of morality. It is clear that from this approach traditional conceptual frameworks for interpreting and assessing moral acts of human beings have to be supplemented by conceptual frameworks that make sense for moral acts of technical artefacts.

Apart from attributing moral agency to technical artefacts there are other options for trying to ascribe moral significance to technical artefacts. For instance, if it is assumed that technical artefacts by themselves somehow embody morally relevant values then it may be argued that technical artefacts are susceptible to moral judgments on the basis of these embodied values. *Prima facie* this may seem to be a less revolutionary option, because it does not involve the attribution of moral agency to technical artefacts. Nevertheless it still appears to be a rather revolutionary move, for what does it mean to claim that a technical artefact may embody (morally relevant) values? Does it imply that a *material* object may embody values? That would surely be a problematic conclusion. This conclusion may be avoided by arguing that a technical artefact is not just a material object. Technical artefacts are material things intentionally made by humans for particular (practical) ends. This conception of technical artefacts ties them intimately to human agency, in particular, to the ends and values of intentional human action. If it is assumed that apart from a material structure human intentions are somehow constitutive for being a technical artefact, then the ends and values of intentional human actions cannot be decoupled from technical artefacts. This is just one way in which the assumption that technical artefacts may embody morally relevant values may be defended.

Given the radical nature of the idea that technical artefacts exhibit, in one way or another, moral agency or that they embody morally relevant values, it is not surprising that this recent discussion about the moral status of technical artefacts has attracted a lot of attention. The ideas have been and still are fiercely defended and severely criticised. This is reflected in the contributions to this book in which a variety of positions is taken by the authors. This variety may be partly due to differences of opinion about what it at stake when dealing with the issue of the moral status of technical artefacts. Be that as it may, all authors share a common concern, namely how to interpret the moral status of technical artefacts, a problem that becomes ever more pressing with the growing influence of technology on modern life.

The first contribution, by Carl Mitcham, places the discussion of agency and artefacts in the context of a long effort to assess the complexities of human making and using. After discussing pre-philosophical appraisals and briefly reviewing the discussion of intentions in recent analytic philosophy, Mitcham distinguishes three waves of contemporary reflection. A first wave is exemplified by the work of Alvin Weinberg and Langdon Winner, who argue that artefacts can extend human political agency. A second wave is led by Bruno Latour, who contests the implicit primacy of the human and argues instead for the primacy of a network in which humans and artefacts behave as ontological equals. A third wave is initiated by Albert Borgmann and by Braden Allenby and Daniel Sarewitz, who reaffirm human ethical agency interacting with agent-like artefacts. Mitcham concludes by discussing to which extent these three waves of discourse might benefit from a greater engagement with the work of Hannah Arendt.

The next three contributions all defend in one way or another the idea that artefacts can be conceptualized in terms of moral agency. According to Lucas Introna the idea that artefacts have, or embody, some level of agency has become more or less generally accepted. In his view, however, there still exist wide disagreements as to what is meant by the agency of artefacts, how it is to be accounted for, and the subsequent moral implications of such agency. Introna suggests that one's account of the agency of artefacts is fundamental to the subsequent discussion of their moral status and implications. He makes a distinction between two different accounts of socio-technical agency: a human-centred account which he calls inter-actional, and a post-human account which he calls intra-actional. Introna then discusses how the intra-actional account posits the social and technical as ontologically inseparable from the start. This implies a 'co-constitutive' account of agency, in which agency is not an attribute of the human or the technical as such but rather the outcome of their intra-actions. Introna illustrates the implications of his approach by analysing plagiarism detection software, and proposes a 'disclosive ethics' for dealing with the moral intra-actions of humans and technologies.

The contribution by Alan Hanson compares two theories of action—methodological individualism and composite agency theory—together with their associated concepts of moral responsibility. While the theories agree that deeds are done by doers, and that moral responsibility for a deed lies with its doer, they differ on the definition of the doer. For methodological individualism, doers are limited to human individuals. Composite agency theory states that most deeds can be done only by humans working in concert with nonhumans and defines a doer as whatever combination of human and nonhuman entities is necessary to accomplish a deed. Hanson then proceeds by addressing the implications of these theories for understanding responsibility. Methodological individualism limits moral responsibility to human individuals, while composite agency theory attributes it to the combination of humans and nonhumans that did the deed. Hanson shows that in the Western world, methodological individualism is rooted in humanistic modernity, while composite agency theory emerges from postmodernity. And he reviews some non-western examples similar to both composite agency theory and methodological individualism.

Peter-Paul Verbeek, as well, defends the view that moral agency needs to be seen as a hybrid phenomenon involving both human and nonhuman elements. He argues that discussions about moral agency and technology are troubled by a severe misunderstanding. Too often, the claim that technologies are involved in moral agency is misread for the claim that technologies are moral agents themselves. As a consequence, much of the discussion focuses on the question whether not only humans but also technologies can have intentionality, freedom, responsibility, and, ultimately, moral agency. From the point of view of Verbeeks 'postphenomenological' theory of moral mediation, though, this discussion remains caught in a dualist paradigm. It locates human beings and technological artefacts in two separate realms; humans being intentional and free, technologies being instrumental and mute. With regard to the question to what extent technologies can be moral agents, the concept of moral mediation makes it possible to investigate how intentionality, freedom, and agency are in fact the result of intricate connections between human beings and technological artefacts. Rather than asking ourselves whether technologies can meet a pre-given criterion of moral agency, we need to re-conceptualize the phenomenon of moral agency itself in order to understand the moral roles of technologies in our daily lives.

After these three defences of approaching technologies in terms of moral agency, Joseph Pitt's contribution presents a defence of the Value Neutrality Thesis with respect to technological artefacts. He argues that technological artefacts do not contain, have or exhibit values. Technological artefacts can be used for various purposes, some praise-worthy and others not. But that does not mean that the artefacts themselves are praiseworthy or not. To the extent that values can be associated with artefacts, it is through the human decision processes that bring them into being. But since there are so many values inherent in those processes it is impossible to identify any particular one that can usefully be said to characterize that artefact.

In their contribution Ibo van de Poel and Peter Kroes argue against the idea of the value neutrality of technology. They start from the observation that in the context of the approach of Value Sensitive Design (VSD), various proposals have been put forward to integrate moral values in technology through design. These proposals presuppose that technology, more in particular technical artefacts, can embody values. Van de Poel and Kroes investigate whether this idea holds water by examining the neutrality thesis about technology: the thesis that technology is neutral with regard to moral values. They introduce two distinctions with regard to values: (1) the distinction between final value (value for its own sake) and instrumental value, and (2) the distinction between intrinsic value (value on its own) and relational or extrinsic value. This leads to four different kinds of values to which the neutrality thesis may refer. They argue that the most interesting version of the neutrality thesis refers to extrinsic final values and provide a number of counterexamples to this version of the neutrality thesis. On the basis of these counter-examples they suggest a general account of when a technology may be said to embody values. This brings them to a distinction between three different kinds of values involved in a design process; intended value (the value intended by the designers), embodied value, and realized value (the value that is realized in actual use).

Philip Brey argues that positions in favour of the view that technological artefacts are or can be moral agents are ultimately lacking because they obscure important differences between human moral agents and technological artefacts. As an alternative, he develops an approach which he calls 'structural ethics', and which does not ascribe moral *agency* to artefacts, but rather moral *roles*. Structural ethics focuses on ethical aspects of social and material networks and arrangements, and their components, which include humans, animals, artefacts, natural objects, and complex structures composed of such entities like organizations. In his view, components of networks that have moral implications are called moral factors. Artefact ethics, then, studies how technological artefacts may have a role as moral factors in various kinds of social and material arrangements as well as across arrangements. Brey argues that his structural ethics and artefact ethics provide a sound alternative to approaches that attribute moral agency to artefacts.

Deborah Johnson and Merel Noorman start from the idea that artefacts, in combination with humans, constitute human action and social practices, including moral actions and practices. Their concern is with the question of what is regarded as a moral agent in these actions and practices. Discourse on artefactual agency and artefactual moral agency seems to draw on three different concepts of agency, they claim. The first has to do with the causal efficacy of artefacts in the production of events and states of affairs. The second can be thought of as acting for or on behalf of another entity; agents are those who perform tasks for others and/or represent others. This concept draws an analogy with humans acting for other humans, but this analogy often blurs the difference between delegation of tasks and delegation of responsibility. And the third concept of agency is distinctively moral and depends on the notion of moral autonomy. Attributions of moral agency to artefacts make sense when they refer to the causal efficacy of artefacts or to the tasks that have been delegated to artefacts by humans. Attempts to extend moral autonomy to artefacts, Johnson and Noorman claim, seem to move from metaphor to status. These attempts claim that humans and machines are analogous and, then, attribute to artefacts the status associated with moral autonomy on the basis of the analogy.

Christian Illies and Anthonie Meijers argue in favour of an active role for artefacts in morality, without introducing radically new moral agency concepts. They develop a tool for the ethical evaluation of artefacts, which they call the 'action scheme'. An action scheme is the repertoire of possible actions available to an agent or a group of agents in a given situation. The action scheme of an agent in a concrete situation is determined by many different parameters, which can be located in physical, intentional, and social frameworks. When artefacts are introduced, they alter an agent's action scheme; new options become available, and some are made more, some less, attractive. The 'action scheme' tool allows designers to analyse and evaluate the effects of artefacts on users in a systematic way; it can show them in what ways artefacts can influence what agents are likely to do. The agent remains responsible for what he or she does. But the designer (and others involved in the creation of artefacts) has a 'second-order responsibility' for changes in the user's action scheme.

Luciano Floridi argues that artificial agents extend the class of entities that can be involved in moral situations, for they can be correctly interpreted as entities that can perform actions with good or evil impact. In his contribution, he analyses the concepts of agent and of artificial agent and then distinguishes between issues concerning their moral behaviour versus issues concerning their responsibility. The conclusion is that there is substantial and important scope, particularly in information ethics, for the concept of moral artificial agents not necessarily exhibiting free will, mental states or responsibility. This complements the more traditional approach, which considers whether artificial agents may have mental states, feelings, emotions and so forth. By focussing directly on "mind-less morality", Floridi shows that it is possible to by-pass such questions as well as other difficulties arising in Artificial Intelligence, in order to tackle some vital issues in contexts where artificial agents are increasingly part of the everyday environment.

The remaining contributions to this book explore the idea to what extent normative evaluations of technologies are possible. Maarten Franssen investigates the ways in which technical artefacts can be subject to normative judgements. When we speak of good saws, poor drills, and so forth, the judgements concern the instrumental value of artefacts: a saw is good as a saw, a drill is poor as a drill. Franssen investigates whether it is also possible to attribute non-instrumental value to artefacts. Can we judge an artefact to be good or bad not in the sense of being an instrumentally good saw or poor drill but being a morally good saw or bad drill? He develops a view of normativity that takes reasons for action or thought as the fundamental notion and that links the value of entities to the existence of reasons to create or promote them in case of positive value or goodness and to eliminate or fight them in case of negative value or badness. He argues that artefacts can be evaluated as bad or good not on the basis of how they are used but on the basis of their design. Additionally, Franssen investigates whether this analysis applies to judgements of artefacts as good just as much as it applies to judgements of artefacts as bad, in order to reveal an asymmetry between the two. And finally he extends his analysis to moral judgements about other artefacts than technical ones, notably works of art.

Finally, Sven Ove Hansson argues that there are substantial similarities in how value statements are applied to chemistry and technology. Both disciplines are subject to negative moral valuations due to the harmful effects of some of their products. In addition, instrumental value statements of a specific type, namely category-specified value statements, are used in both areas. Examples are "a bad engine" and "a good stabilizer". In both cases, this usage is based on functional descriptions that relate to the design component of the respective discipline. However, there are also important differences in how such value statements are applied in chemistry and in technology. Hansson investigates these similarities and differences, and concludes that additional studies along these lines can contribute to our understanding of values and technology.

Given the diversity of opinions defended in this book it may be likened to an intellectual smörgåsbord for anyone interested in the moral significance of technology and technical artefacts. Whether one prefers to reserve moral agency or moral values exclusively to the domain of human beings or one wants to include technical

artefacts in one way or another as well in the domain of moral agents or morally valuable things, the contributions to this book offer a plethora of arguments and counterarguments that have to be taken into consideration by anyone who is interested in the moral significance of technical artefacts. However, the topics discussed in this book are not only relevant to philosophers of technology interested in the moral status of technology. There are many interesting connections to other philosophical sub-disciplines and even to disciplines outside philosophy. Obviously the radical nature of the thesis of the moral agency of technical artefacts raises all kinds of meta-ethical issues. But there are also intimate ties, although often not explicitly pointed out in the various contributions, to discussions in philosophy about distributed agency, distributed responsibility, the extended mind and collective intentionality. Some of the topics addressed in this book are directly relevant to fields outside philosophy, like design methodology and design research. Given these connections to other fields of study it is to be hoped that the study of the moral status of technical artefacts will not stay confined to the philosophy of technology. There may be a lot to be gained, for the philosophy of technology and for these other disciplines, by confronting together the problem of the moral status of technical artefacts.

## Reference

Heisenberg, W. (1958). The representation of nature in contemporary physics. *Daedalus, 87*(3), 95–108.

# Chapter 2
# Agency in Humans and in Artifacts: A Contested Discourse

**Carl Mitcham**

**Abstract** Philosophical discourse on agency and artifacts is part of a long effort to assess the complexities of human making and using. Appreciation of some historico-philosophical aspects of the discussion begins with a sketch of pre-philosophical appraisals; then, given common assumptions about the importance of intentionality in agency, ventures a brief review of the debate about intentions in recent analytic philosophy. Against this dual background, contemporary reflection more specifically on agency and artifacts is distinguished into three waves. A first wave is exemplified by the work of Alvin Weinberg and Langdon Winner, both of whom argue that artifacts can extend human political agency. A second wave is led by Bruno Latour, who contests the implicit primacy of the human and argues instead for the primacy of a network in which humans and artifacts behave as ontological equals. A third wave is initiated in critical works by Albert Borgmann and by Braden Allenby and Daniel Sarewitz, who argue for deploying first and second wave insights to reaffirm human ethical agency interacting with agent-like artifacts. A conclusion considers how these three waves of discourse might benefit from engagement with such historico-philosophical studies of human agency as found in the work of Hannah Arendt.

From the earliest periods of thinking about physical objects, human beings have experienced tensions between themselves and the material world, including paradoxically their makings. One can easily imagine a prehistoric knapper dependent on his stone tools and thus thinking of them as good, but accidentally cutting himself while skinning an animal and complaining, "This thing has an evil spirit." The skin fashioned into a garment for warmth can also inhibit the chase of quarry, at which point it would have had to be temporarily discarded for "holding back" a hunter. On the one side,

C. Mitcham (✉)
Colorado School of Mines, Golden, CO, USA
e-mail: cmitcham@mines.edu

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_2,

humans recognize that artifacts extend agency (making it more effective); on another, artifacts can undercut human activity (distorting it or causing harm). Intimations of such agency-artifact tensions can be found in archaic myth and legend; think only of the stories of Cain and Able or of Daedalus and Icarus.

The tension between human agency and artifacts echoes others between humans and their bodies and among humans themselves. Although we are and identify with our bodies, bodies do not always behave as we would like or intend. I want to run faster; I did not intend to get sick. Aristotelian *akrasia* (*Nicomachean Ethics* VII) and the Pauline *fomes* of sin (Romans 7; cf. Thomas, *Summa theologiae* I-II, q. 91, a. 6) reference related experiences at the explicitly moral level. How can I become that person I aspire to be, given the occasional resistance of my body? But how could I aspire to be anything at all without a body, which tells me something about who I am?

With regard to interactions among humans, questions arise that are more political in character: How can I act together with others, when we often fail to share common goals? The problematics of human interactions are further defined by what in law is known as the principal-agent dilemma: How can a principal (agent) who delegates to an (assistive) agent be sure that the assistant, who knows the particulars of an activity better than the principal, will truly act in the principal's interest? Another issue involves collective agency and responsibility, as when a corporation or community acts and is rewarded or punished as a whole. Human agency is never simple but includes the experience of struggling with oneself, others, and one's ostensible tools.

In conjunction with the rise of modern technology and engineering, this tension or struggle has taken new forms. Compare, for instance, celebrations of artifacts as the products of moral agency in Enlightenment praise for the practical arts, interpreted as growing out of and constituting a new and higher level of cultural achievement versus Romantic worries about autonomous makings in stories such as Johann Wolfgang von Goethe's "Der Zauberlehrling" (1797) and Mary Shelley's *Frankenstein* (1818). Across the nineteenth and twentieth centuries and into the twenty-first this salvation-damnation dialectic was manifested in relation to artifacts as diverse as steam engines, electric dynamos, nuclear weapons, synthetic chemistry, computers, digital media, and nanotechnology. A *Time* magazine cover for "The Wireless Issues" (August 27, 2012) proclaimed "Ten Ways Your Phone Is Changing the World" and then went on to identify both positive and negative changes. Technologies clearly seem to be doing things to us as well as for us.

Discussion since the 1800s can be associated not solely with the increased density of lifeworld artifice but also with new affirmations of human agency and the simultaneous positing of countervailing forces. As if to prove that human beings can and should take control of and transform the world—that humans are more properly understood as autonomous agents than had ever previously been thought (*vide* Immanuel Kant's "Beantwortung der Frage: Was ist Aufklärung?" of 1784)—captains of industry created mass production economies and new technologies of transport and communication that transcended all natural or traditional boundaries. Conceived as having god-like agency humans became god-like transformers of the lifeworld. At the same time, questions arose about whether agency, as the capacity to make choices and impose them

on the world, was properly assigned to individuals. G.W.F. Hegel, for instance, who described the principal-agent problem in the peculiarly fraught terms of a master–slave dialectic, argued that true agency belonged to a historically acting *Geist*; for Karl Marx the historically important agent was a social class. In economics and sociology individual human agency is often argued to be subordinate to larger socio-economic forces and socio-cultural structures. By contrast, Sigmund Freud, evolutionary biologists, and now neuroscientists point to non-conscious psychological and physiological factors operative in alleged individual decision making.

This discourse has taken a new and distinctive turn with discussions about the possible transfer of moral agency to artifacts, in some cases specifically designed as such. Increasingly inserted into these contemporary interpretations of human agency have been efforts to analyze artifacts themselves as agents (see, e.g., Adam 2005; Harbers 2005; Johnson and Powers 2005; Verbeek 2005, 2011; Floridi and Sanders 2004). These efforts simultaneously reflect and set the stage for transformations in engineering, corporate organization, and lifeworld experience as well as in philosophy. Within philosophy questions are asked that suggest the need to rethink the nature and meaning of agency both in the present and in the past, not to mention the future. Against this condensed historico-philosophical background, what follows is simply a set of further reflective notes on the emerging discourse concerning relations between moral agency and artifacts.

## 2.1 Intentions, Ethics, and Artifacts

Central to agency appears to be some notion of intention. How can one be an agent and be morally accountable for an action without intending it. Folk consequentialists as well as folk deontologists assume the intentions of agents as fundamental to the ethical assessment of human actions. Indeed, intentionality, as implicating an inner life that includes the ability to imagine objects and states of affairs, both past and prospective, together with willing a state of affairs, seems to be a *sine qua non* of deep moral conduct. Yet as Joseph Shaw has observed, "The use of intention in ethics has been the subject of intense debate for many years, but no consensus has emerged over whether intention is morally relevant, or even how it should be understood" (Shaw 2006, p. 187). The common parlance that relies on intention when assessing action is questioned in analytic philosophy.

Efforts to elucidate the precise character of intentions have involved philosophy of mind and action, psychology, and most recently even neurophysiology. Shaw, for example, draws multiple distinctions among intention and causation, desire, motive, moral responsibility, and even what is done in intentional action. This conceptual analysis introduces complexities not amenable to easy summary, a few relevant contributions deserve mention.

Some notion of intention appears crucial to any account of human action. Only intentional behavior constitutes action in the most serious sense. As Ludwig Wittgenstein (1953) famously observed, there is a difference between my arm

going up and me (as an agent) raising my arm. Subsequent philosophical analyses by G. E. M. Anscombe (1957) distinguished between desire, belief, and intention as components of agency. Having intentions has been argued to mean that agents both desire to do something and believe that they will do it.

This view has been developed and criticized by both Donald Davidson (1980) and Michael Bratman (1987). According to Davidson, intentions are best described as evaluative judgments. According to Bratman, the most effective way for humans living in association with others to become effective agents is to have plans, one element of which is intentions. Plans (which he also describes as "intentions writ large") are composed of beliefs, desires, and intentions (writ small). This BDI (belief-desire-intention) model has been implemented in computer programs. Beliefs, constituting the information state of an agent, are stored in a data base. Desires or goals are programmed into devices so that they transform specific inputs into specified outputs. Intentions then select from a repertoire of plans or movement sequences in order to achieve the "desired" input–output function. In a computer controlled HVAC system a program to maintain a certain ambient temperature could, when the temperature rises too high, select the plan of shutting off the heater or turning on the cooler, depending on input regarding the current functional state of system. The BDI model is, however, dependent on the prior formation of a suite of action plans properly linked to appropriate belief states. As Bratman subsequently restates his planning theory,

> intentions are characteristically elements of larger, partial plans of action, and these plans play basic coordinating, organizing roles at a time and over time. Associated with these roles are distinctive rational pressures on intentions for consistency and coherence at a time, and stability over time. (Bratman 2007, p. 5)

More broadly, Bratman contends that strong human agency is reflective, planful, and temporally extended—an analysis that might be interpreted as applicable to machine agency.

A further distinction argued by John Searle (1983) describes two ways mind can relate to the world, both of which are described in terms of intentionality. In the one type of intention, manifest in acts of perception and knowing, people seek to alter their minds so that ideas properly reflect or fit with what is in the world. This is the intentionality (sometimes spelled "intension") examined at length by philosophers such as Daniel Dennett (1987). In the other, in acts of volition or willing, people try to make the world fit an idea in the mind. The two types of intentionality are described in terms of "direction of fit": mind to world or world to mind. (Searle's distinction was, as he admits, anticipated by Anscombe; see, e.g., Searle 1985, p. 3; and Anscombe 1957, §32. It is also, Anscombe might well argue, simply a restatement of the scholastic understanding of truth; see Thomas, *Summa theologiae* I, q.16, a. 1.) Although Searle himself explicitly denies the possibility of machines having intentions of either type, it remains possible to distinguish between devices designed to sense states of affairs in the world and devices designed to alter or manage states of affairs.

In psychology intention is defined as the mental state that obtains when a person has a number of options, chooses to follow one of them (especially when non-dominant), and focuses attention on this option. Intentions and intentionality in the

psychological sense are at the basis of social interaction, since it is the ability of one person to appreciate the intentions of another that gives human interaction its distinctive character, and is central to analyses of criminal responsibility (Malle et al. 2001).

In all such senses intentions are intimately involved with agency and ethics, yet in most it is difficult to attribute intentions in anything more than a metaphorical sense to artifacts. Only in the most advanced form of artificial intelligence is it reasonable to describe artifacts as having (rather than simply embodying—or, perhaps, *endigitalizing*) plans or options from which they select some particular functional state. Indeed, the most common argument against conceiving artifacts in terms of agency is that artifacts cannot have intentions. Artifacts do not have minds, an inner life, or a theory of mind, although they may represent or model minds, inner lives, or theories of mind. To alter their behavior one must re-program them; it is not possible to teach them to consider alternatives. To alter the operation of artifacts in any fundamental way one must open them up and change their insides. In extreme cases, of course, human behavioral changes can also require altering an anatomy or physiology. But with humans there are also instances in which they can be invited to change themselves, to take it into and upon themselves to alter their behavior, even by opening up their own insides or their minds.

From the perspective of conceptual analysis there is thus some common sense truth to the negative view that artifacts, even complex ones, are not agents, because they have no inner lives that they themselves can change. At the same time, outside analytic philosophy the idea that artifacts do not have intentions is at odds with experiences in which artifacts do appear to have "plans" for their use, plans of which their users must be aware in order to interact with them effectively. Although we quickly dismiss the idea as just a poetic way of thinking, we even say at times that machines anticipate our wants and needs, as when the house appears welcoming or the car is ready to take us somewhere. As radical expressions of this perspective, see Stewart Brand's *How Buildings Learn* (1994) and Kevin Kelly's *What Technology Wants* (2010). As Kelly describes his experience growing up in

> New Jersey in the 1950s and 1960s, I was surrounded by technology. But until I was 10, my family had no television, and when it did arrive [I noticed how the] TV had a remarkable ability to beckon people at specific times and then hold them enthralled for hours…. They obeyed. I noticed that other bossy technologies, such as the car, also seemed to be able to get people to serve them, and to prod them to acquire and use still more technologies (freeways, drive-in theaters, fast food)…. As a teenager, I was having trouble hearing my own voice and it seemed to me my friends' true voices were being drowned out by the loud conversations technology was having with itself. (Kelly 2010, p. 2)

Beyond artifacts that we often experience as "wanting" us to use them in specific ways or otherwise influence and delimit our actions, robot pets from the late 1990s Tamagochi and Furby to a host of increasingly sophisticated devices are designed around the conceit of intention possessing artifice. For divergent appreciations of this phenomenon, see David Levy's *Love and Sex with Robots* (2007), Sherry Turkle's *Alone Together* (2011), and the movie "Robot and Frank" (2012).

Responding to the rejection-of-intentions argument and struggling to reflect the kind of experience described by Kelly are efforts to attribute intentions to artifacts,

to discover and analyze ways in which artifacts may be said to have intentions or agency designed or embedded in them. Whenever humans experience problematic phenomena there are always at least two possible philosophical responses: argue that the phenomena are illusory or that there is more going on than has been previously appreciated. Although absent primary agency and intentions of the kind human beings possess, perhaps artifacts possess what might be termed secondary or imposed agency and intention. The story of this philosophical effort to think intentions in artifacts can be described in terms of three waves.

## 2.2 Artifacts with Secondary Agency

With regard to the idea of artifacts exhibiting secondary agency, there are two classic, first wave texts: Alvin Weinberg's "Can Technology Replace Social Engineering?" (1966) and Langdon Winner's "Do Artifacts Have Politics?" (1980).

At the time of articulating his argument, nuclear physicist, engineer, and self-described "technological fixer" Weinberg was director (1955–1973) of the Oak Ridge National Laboratory. He began by distinguishing between technological and social problems. The former are exemplified by the tasks of building an atomic bomb, creating radar, or launching Earth-orbiting satellites; the latter by controlling world population and protecting the environment. "In view of the simplicity of technological engineering, and the complexity of social engineering," he asked, "to what extent can social problems be circumvented by reducing them to technological problems?" (Weinberg 1966, p. 5). He answered by citing two cases he considered dramatic successes: the reduction of poverty by increased industrial production instead of social revolution and the limitation of warfare by nuclear weapons. In the future, he proposed that technological birth control will limit overpopulation and that cheap energy from non-polluting nuclear power would protect the environment. "The Technological Fix accepts man's intrinsic shortcomings and circumvents them or capitalizes on them for socially useful ends." Yet Weinberg admitted that "technological solutions to social problems tend to be incomplete and metastable, to replace one social problem with another" (Weinberg 1966, p. 8).

Nowhere is this displacement problem more obvious than with regard to automotive safety. Weinberg favorably referenced Ralph Nader's argument for technological remedies to the problem of increased harms from automobile accidents (Nader 1965). What he did not acknowledge is that according to Nader's analysis the implementation of technological fixes depends on a demand by citizens or the government. They will not come about of themselves through simple technological progress, engineering design, or free market competition. The intentions of making a profit and defending the status quo are so strongly invested with those who control technologies that there are often serious social and technologically embedded resistances to change. Indeed, in a strange way Weinberg's emphasis, like that of Nader's, remains focused on human and social agency: How to get people to recognize their own best interests and then generate intentions to use technologies to more

effectively realize these interests. Multiple efforts to affect technological fixes have been explored by others (Rosner 2004) and defended against critics of what are sometimes disparagingly termed "technological shortcuts" (Etzioni 1968; Etzioni and Remp 1973).

A decade and a half after Weinberg, political theorist and science-technology-society studies scholar Winner shifted the argument. For Winner technological fixes were not so much agents for the realization of human interests, but ways in which technology unbeknown to users, influenced their lives. Citizens need to wake up not so much to their interests, which they can then realize through technology, but to how their interests are being frustrated by technologies, which in some cases manifest counter-interests. For Winner technologies can have political implications—something related to agency and intentions—all their own.

Winner outlined two ways that artifacts can embody politics: intentionally and unintentionally. The Long Island overpasses of New York City planner Robert Moses, the Parisian boulevards of Georges-Eugène Haussmann (a French counterpart of Moses from the previous century), and the molding machines of American industrialist Cyrus McCormick were all designed to "enhance the power, authority, and privilege of some over others" (Winner 1980, p. 125). Hidden within technological artifacts, publicly justified on technical rather than political grounds, were secret political decisions and intentions. Moses wanted to keep poor African Americans, who would have had to use buses that could not negotiate his overpasses, from certain suburban areas; Haussmann intended to prohibit the barricades of another Paris commune; and McCormick aimed to limit the power of labor unions. David Noble's analysis (1977) of machine tool automation provides a further vivid example of the embodying of anti-union politics in engineering designs. In each instance technologies were created to achieve the intentions of their designers or promoters.

Other artifacts, however, manifest political influences beyond any conscious intent of their engineers. Buildings that have grand steps to their entrances in order to establish a sense of wealth (banks) or power (court houses) unintentionally limit access by persons with mobility handicaps. The mechanical tomato harvester, engineered to reduce manual labor, because of its cost, unintentionally promoted economic relationships that put small family farms out of business. With regard to how some artifacts can have their own politics, Winner further distinguished two possibilities. In one (strong) case, a technical system "*requires* the creation and maintenance of a particular set of social conditions as the operating environment of that system." In another (weaker) case, a "technology is strongly *compatible with*, but does not strictly require, social and political relationships of a particular shape" (Winner 1980, p. 130, Winner's emphasis).

Exemplifying the first, Winner cited an analysis of nuclear power plants by Lewis Mumford (1964). Nuclear power systems seem to demand some degree of authoritarian or centralized control in the social orders in which they operate. Exemplifying the second case, although not cited by Winner, are the irrigation systems of Karl Wittfogel (1957), who argued that the hydraulic engineering of landscapes tends strongly to be associated with centralized and authoritarian social

orders ("oriental despotism"). As both examples suggest, the idea of artifact agency implicates notions of technological determinism.

Bruce Bimber in "The Three Faces of Technological Determinism" (1994) refines the strong versus weak versions of technological determinism into normative, nomological, and unintended consequences versions of how artifacts can "intentionally" determine social orders. Normatively, societies may project cultural and political meanings onto technologies; across the twentieth century, for instance, normative value was regularly ascribed to technical (and economic) efficiency (see also Alexander 2008). Nomologically, and internally, one kind of technology may necessitate or promote another; it is difficult to image the four-cylinder internal combustion engine being developed before the single cylinder engine or the internal combustion engine before the steam engine. Another term for nomological determinism is "path dependency." Analyses emphasizing inner determinations in the evolutions of technology are present in authors as diverse as Gilbert Simondon (1958) and Brian Arthur (2009). Finally, technologies or artifacts often have unintended consequences either as side effects or as second, third, n-order effects (see Averill 2005). In response to the unintended consequences thesis, however, Michel de Certeau (1980) and Andrew Feenberg (1999), among others, have argued that human beings are as creative in adapting technologies to new intentions as technologies are in foisting their own secondary intentions off on humans. Just as the agency of one person can provoke the agency of another, so can the secondary agencies of artifacts promote or stimulate active human responses.

Two further examples of arguments to the effect that artifacts can, under certain circumstances, exhibit secondary agency are present in works by Anders and Eartherly (1961) and Ivan Illich (1973). Austrian philosopher Anders carried on a series of exchanges with Claude Eartherly, who flew the reconnaissance plane on the Hiroshima bombing mission and apparently came to feel culpable for the (unintentional—at least on his part, personally) results of his actions. In the course of these exchanges Anders attempted to extend deontologism from persons to technological artifacts. In the eighteenth century Kant had formulated the fundamental principle of morals as "Act only according to that maxim that you can at the same time will should become a universal law." To extend this principle into the engineered world, Anders reformulated it as: "Have and use only those things, the inherent maxims of which could become your own maxims and thus the maxims of a general law" (Anders and Eartherly 1961, p. 18). Insofar as the inherent maxim of a nuclear weapon is mass destruction, it cannot be used morally.

In a similar spirit Illich, an anarchist critic of technoscientific development, attempted to identify the inherent maxims of certain types of technology. Going beyond Illich's own phrasing, the tool as a functional entity is dependent on two kinds of human inputs or engagements: energy from a user's body muscles, and formal guidance from hand-eye coordination (think of a hammer). By contrast, machines rely on non-human energy inputs (the internal combustion engine or electrical power) and only directly on formal guidance from humans (as in the driver of a car). When the immediate formal guidance for a machine itself becomes a computer program (as in something so simple as an HVAC system), still more

separation is introduced into the artifact-human relationship. These different levels of separation—or alienation—constitute general maxims to be reflected on in the human use of technologies, just as when collaborating with other people one learns to pay attention to and take into account their motives or intentions.

## 2.3 Artifacts as Delegated Agents

For Anders, Mumford, Weinberg, Illich, and Winner artifacts may function as agents only in some secondary sense. The secondary sense may have positive features (Weinberg) or challenging ones (Winner), but all agree that agency remains or ought to remain primarily with the human designers and users. However, in a second wave of reflection on agency in artifacts in the 1980s a small group of sociologists of science, technology, and society (STS) studies proposed a new approach called Actor Network Theory (ANT) that rejected any hard distinction between primary agency in humans and secondary agency in artifacts. The leading figures of this new approach were Bruno Latour, Michel Callon, and John Law (all contributors to Law 1991; and to Bijker and Law 1992)—but especially Latour. For Latour the distinction between agent actors and their artifacts is to be replaced by a concept of "actants" that applies equally to both.

The concept of an actant, was initially imported into STS studies from the narrative theory of Algirdas Julien Greimas (Latour 1984, 1987). In narrative theory actants are paired actors dependent on the pairing for their dramatic features; examples include hero and villain or lover and beloved. For Latour an actant is simply anything that "acts or shifts actions" (Akrich and Latour 1992, p. 259), action being further defined by the results of a set of tests or practical interrogations—that is, by whatever results an actant actually brings about under specified conditions. Take a vial of X and introduce chemical or microbe Y. If Y alters X then Y is an actant, and can be described as an agent precisely in terms of the changes it introduces. The act of introduction can be ignored. Note how even in pre-Latourian chemistry, scientists used the term "re-agent" (without the hyphen) to reference "any substance added to a system in order to bring about a re-action" (again, without the hyphen). Latour thus adapts the language of chemistry and valorizes it for STS studies usage—although in chemistry the "re" importantly suggests a secondary agency or action.

Once in place, the actant concept can be used to transform social network analysis, which has its own varied history reaching (network-like) back to the 1800s. For instance, the classic sociologist Ferdinand Tönnies (1855–1936) argued that personal ties could be either direct and personal (as in traditional *Gemeinschaften*-communities) or impersonal and formal (in *Gesellschaften*-societies), the latter characterizable in terms of individual node-based networks (although he did not use these exact words). In the 1930s social psychologists such as Elton Mayo (1880–1949) and J.L. Moreno (1889–1974) analyzed social interactions in small groups to understand the workings of industrial bureaucratic organizations. In the 1950s and following anthropologists deployed social network analysis (in, e.g., kinship

studies) to describe non-European social orders. The 1970s witnessed the emergence of distinctive schools of social relationships using social network analysis to describe political and economic activities, an approach that has became ever more relevant with the innovations of social networking media. In all these instances, however, the actors remain human agents with intentions that could nevertheless be transformed in multiple ways. Expanding the notion of what might constitute a node in these social actor networks leads to ANT, which is more properly termed Actant Network Theory.

As in social network analysis, ANT argues that actants in a network become agents of a particular sort by virtue of their relationships with one another. For heuristic purposes, it assumes that nothing lies outside network relations, and suggests there is no difference in kind in the abilities of humans, animals, or artifacts to act or shift action. "Society and technology are not two ontologically distinct entities but more like phases of the same essential action" (Latour 1991, p. 129). Instead of ontologically distinct entities, technology is simply "society made durable" insofar as artifacts can be characterized in terms of "scripts" that inventors and engineers write into them regarding their uses and interactions with both humans and non-humans (Akrich 1992, p. 207ff.). As soon as a human engages with some established actant network it is caught up in an existing web of relationships into which it can also introduce change by its very engagement. In a further elaboration of his ontological thesis, Latour argues "the impossibility of having an artifact that does not incorporate social relations, as well as the impossibility of defining social structures without accounting for the large role played in them by nonhumans" (Latour 1999a, p. 212).

Parallel to his effective rejection of any distinction between primary and secondary agency Latour develops the notion of delegated agency, which would seem to recreate the difference. In an analysis of four different types of technological mediation—interference, composition, folding of time and space, and crossing the boundary between signs and things—Latour (1992) observed how, in the fourth instance, things become signs and signs things. Take the case of a speed bump: Is it not a sign ("Slow Down") become a thing (a piece of raised concrete on a street that causes what it signifies)? An action (the enforcement of a speed limit) has been delegated to an artifact. The apparent reason why this is not just secondary agency: for Latour all agency is delegated—in this case, either to a policeman or to a speed bump. Even the delegation, by the traffic safety manager or the engineer, has to have been delegated: by the law or bureaucracy or some written instructions. An actant-network is ultimately a network of delegated actants: It's delegation all the way down.

Can delegated agency be agency? Is delegated morality true morality? The question cannot help but suggest issues related to representation, another key concept in the Latour lexicon (see Latour 1999b) that has been further examined by Mark Brown (2009). According to Brown, scientific or epistemological representation is properly complemented by political or governmental representation. But there are two modes of political representation: trustee and delegate. In the trustee mode, representatives are asked to stand in for appropriately designated groups by exercising or acting on their own best judgments. The trustee representative functions like a

physician who knows more about patients than patients know about themselves. In the delegate mode, representatives are not at liberty to act on their own; instead they must act only as their constituents would. Delegates are simply guns for hire. (The epistemological analogates might be instrumentalist and realist interpretations of concepts. In instrumentalism, concepts are modestly independent of the phenomena they represent; in realism, concepts are more closely tied to their phenomena).

Considering the complexities of representation, it is not clear that delegate/realist theories allow for agency in any strong sense. For instance, according to the delegate theory, representatives in parliament should simply vote the way those they represent would vote. By contrast, trustee/instrumentalist theories of representation would allow representatives in parliament to make decisions and act on the basis of expertise and knowledge developed independently of those they represent. In this respect, when Latour writes of morality delegated to artifacts, it would seem to undermine the notion of artifacts as being able to act morally.

Another objection to ANT is that it simultaneously anthropomorphizes artifacts and objectifies humans (Collins and Yearley 1992). Insofar as humans are posited as co-constituted by nonhumans, it is unclear to what extent humans can be responsible for their moral character or decisions. "How could someone be praised or blamed for an action or intention if these are constituted by a continuously shifting network of associations?" (Mitcham and Waelbers 2010, p. 379).

In an extension of the objection to equating action in humans and nonhumans, Harry Collins and Martin Kusch (1998) developed an insightful analysis that distinguishes between polimorphic and mimeomorphic actions. Polimorphic actions can look differently in different instances and yet be the same action; mimeomorphic actions look alike whenever they occur. Insulting someone can take many forms and not always look like the same action even though it is, since it can be done with different words or even without words. Stirring the pot is an action that looks pretty much the same whenever it takes place. On their argument, humans can perform both kinds of actions, but machines or artifacts only the second.

This sociological theory of action is developed independently but not in ignorance of discussions in the philosophy of action and in fact has much in common with a hermeneutic theory of action. Hubert Dreyfus's argument regarding "what computers can't do" (1972), for instance, relies on a related analysis. As a sociological action theorists, Collins and Kusch are concerned not with actions as individual events (action tokens) but actions as repeatable events (action types). Additionally, their analysis distinguishes between action types that are possible because of the existence of a social order and those that are not.

> In British society in the 1990s we can go to the cinema, drive to work, play darts, supervise a student, take out a mortgage, and so on. If we were members of the Azande society …, we could do none of these. We could, however, accuse someone of being a witch, prepare *benge*, consult the poison oracle, and invoke spirits; in current British society we can do none of these [as a result of] the differing social and conceptual structure of life in the two societies. (Collins and Kusch 1998, p. 7)

From this perspective, agency—whether delegated or not—is a function of or dependent on culture, which is also informed by technology.

## 2.4 Artifacts and Cultures

The Wittfogel thesis (as mentioned in relation to political theorist Winner) calls attention to how big histories often construct narratives around mutual interactions between artifacts and culture. Anthropologists traditionally periodize the long sweep of human cultures into stone, bronze, iron, and related ages. More recently, Thomas Hughes (2004) crafted a succinct overview of modern technology in terms of four periods of cultural interaction. Hughes thus provides a convenient framework for thinking about technology and culture can interact to influence both human and artifact agency.

In the first period, which emerged from Christian intentions, technological artifacts were envisioned as a new or second creation. There was nevertheless a tension between visions of a new creation that would be enclosed within the old (as in the theory of Hugh of St. Victor) or supercede and replace it (as in Francis Bacon). In North America this tension was echoed in a tension between visions of paradise regained (J.A. Etzler) and pastoralism (Thomas Jefferson). In both, human intention remained paramount.

In a second period, the new industrial revolution of Thomas Edison and Henry Ford aspired to transfer the order of the machine into the human world. The non-human machine was taken as a model for human order and organization. There was an attempt to discover a pattern in artifice that should become an intention for humans. In so doing, the mechanical pattern itself tended to be treated as intention-like. Such figures as Oswald Spengler, Lewis Mumford, Werner Sombart, and Walter Rathenau analyzed the tensions between old and new, organic and mechanical, human and machine "intentionalities" or patterns of behavior that gave rise to what was often termed "machine culture."

This projection of machine orders into the intentional proliferation of artifacts led to the creation of an technological complexity unprecedented in history: the rise of systems that appeared to extend beyond traditional forms of human control. The result was an explicit conceptualization of machines in the new science of cybernetics as having their own special kinds of (artificial) intentions and a corresponding reconceptualization of human intentions in machine control terms. The post-World War II era of cybernetics and systems (Norbert Wiener and others) bled into the information revolution that spanned engineering, biology, sociology, and politics and then opened up into the world of computers, the internet, and biotechnology. Traditional versions of intentionality were repeatedly reframed in information-theoretical and computer-related terms (via artificial intelligence, cognitive psychology, and eventually genetics become genomics). What Hughes called "technological values" "infused" art, music, architecture, and the manifold of culture beyond anything previously imagined in the machine age.

At the same time the fate of the material base in nature re-conceived as environment—or that which surrounds humans—became a theme of concern. Concepts of environmental crisis and sustainability sponsored the emergence of a new intention, one oriented toward the creation of alternative eco-artifacts, which themselves have

often been described in intention-like terms as more friendly toward both the human and non-human worlds.

Each of these four different artifact-culture interactions manifest dependencies on human intentionality in creation or in use—while themselves reforming human intentions in an often intention-like if unintended manner. On the basis of such a history one might adapt Baruch Spinoza and postulate a non-reductionist, dual-aspect theory of artifice as both independent and dependent. One aspect would be the macro-independent theory of autonomous technology as formulated by Jacques Ellul (1954; and revisited in Jerónimo et al. 2012); another would be the micro-dependencies articulated by social deconstructivism (e.g., in actor network theory). Just as the mental and the physical are argued to be different aspects of one substance, so autonomy and social construction can be postulated as double attributes of one underlying reality; the larger truth is not either/or but both/and, depending in part on levels of analysis. Of course, just as with mind-brain dualism, there is the problem of analyzing clearly relationships between the two aspects.

One ethics-related implication of any dual aspect theory is an obligation to take both aspects into account. The first- and second-wave reflections regarding agency and artifacts emphasized in the human-technology relationship ontological difference and ontological sameness, respectively. From the perspective of the former, humans need to recognize the ways artifacts differ from and often oppose human agency, in order to better exercise their own agency when using technology. In the latter, humans need to recognize a common bond between themselves and artifacts, to overcome what Bruce Mazlish (1993) termed the "fourth discontinuity" and accept the "co-evolution of humans and machines." In what may be called a third wave of reflection on agency and artifacts, there emerges an effort to synthesize the two perspectives as part of a normative project for living more consciously with artifacts.

This third-wave reflection is exemplified in Albert Borgmann's *Real American Ethics* (2006) and in Braden Allenby and Daniel Sarewitz's *The Techno-Human Condition* (2011). Borgmann's contribution to this third-wave reflection can actually be traced back more than three decades (see, e.g., Borgmann 1984). But the book on "real ethics"—as distinct from theoretical and applied ethics—restates Borgmann's argument in what he calls the "Churchill principle," which takes its name from Winston Churchill's remark, apropos debate about rebuilding the parliament building after it was bombed in World War II, "We shape our buildings, thereafter they shape us." We socially construct a network of artifacts (as emphasized in second-wave reflections) which also influence us (emphasized in first-wave reflections). Because of both aspects, we need to think normatively about who we are and who we want to become. The challenge of agency and artifacts is to reflect on the good.

> The ways we are shaped by what we have built are neither neutral nor forcible, and since we have always assumed that public and common structures have to be one or the other, the intermediate force of our building has remained invisible to us, and that has allowed us to ignore the crucial point: We are always and already engaged in drawing the outlines of a common way of life, and we have to take responsibility for this fact and ask whether it is a good life, a decent life, or a lamentable life that we have outlined for ourselves [with our artifacts]. (Borgmann 2005, p. 6)

More than any other philosopher of technology Borgmann does not rest with recognition of the two dimensions of artifice. Instead, he argues repeatedly and progressively not just for increased consciousness of our condition but for recognition, defense, and pursuit of the foundational goods of social justice (appreciating the reality of others) and environmental stewardship (protecting the reality of nature). In the world of intensified artifice such abstract goods can be engaged through what he terms "focal things and practices" of the "culture of the table" and communal celebrations of voluntary simplicity (Borgmann 2005, p. 160). "Our most recent technological culture, due to its highly mediated and virtual character, brings the immediacy and actuality of the table and the meal, of family and friends into relief" (Borgmann 2005, p. 199). In the presence of the glamor of commodious devices and consumption, Borgmann invites and encourages us to pursue (as agents) and to construct (in the secondary agencies of artifacts) a more noble material culture and associated way of life than is presently the case.

With less sobriety but no less intensity, Allenby and Sarewitz likewise present a vision of what it means to be good in the presence of the multiple agencies of humans and artifacts. Initially they simply observe that humans have for thousands of years been co-evolving with their technologies, insofar as such activities as tool making and meat eating mutually interacted with brain development and social complexity. What is different now is that we have moved beyond external technological interventions to transform ourselves from the inside out—even as we remake the Earth system itself. Coping with this new reality, they argue, means re-conceiving relationships between technology and nature, recognizing the power and attraction of technological fixes while cultivating a new humility in the absence of certainty about Enlightenment ideals.

In their central analytic proposal Allenby and Sarewitz argue for distinguishing three types of technological cause-effect relationships. Level I causal relationships are straightforward technological fixes. But such relationships are often embedded in social networks that give rise of Level II relationships. The airplane is a Level I fix to the problem of fast long distance transport; but air transport depends on Level II airport construction, air-traffic control systems, financial investments, commercial management, and government safety regulations. Moreover, Level II air transport systems have Level III effects on the global environment. It is relatively easy to understand and depend on Level I relationships, less so on Level II relationships, and still less so on Level III relationships. The basic problem in the agency and artifacts world is that "We inhabit Level III, but we act as if we live on Level II, and we work with Level I tools" (Allenby and Sarewitz 2011, p. 161). Beyond the identification of this condition of incompatibles, however, the authors advance a case for humility and a suite of related normative responses that run from giving up the quest for definitive "solutions," relying on pluralism over expertise, playing with scenarios, increasing the frequency of decision making, questioning predictions, promoting continuous learning, and more (Allenby and Sarewitz 2011, pp. 162–169). Although their focus is more procedural and less substantive than Borgmann's, Allenby and Sarewitz also advance the agency-artifice discourse into a normative possibility

space. Simple consciousness of the agency of artifacts, however this is conceptualized, is not enough; it must be used to enhance distinctly human agency.

## 2.5 Questioning Conclusions

What conclusions emerge from these episodic notes on attempts to think or not agency in artifacts? Some efforts clearly exhibit their own artifice as academic parlor games of rhetorical originality that have failed to appreciate previous analytic achievements. Others are deep efforts to thoughtfully engage the techno-lifeworld in which we all now live and move and have our being. Yet one cannot help but suspect there are other sources that could also be fruitfully placed in dialogue. To test especially the last hypothesis, consider the analysis of Hannah Arendt's *The Human Condition* (1958), to which none of the works examined makes reference. In her effort to think the historical transformations that took place across the first half of the twentieth century, Arendt distinguished three basic types of human activity: labor, work, and action.

First, labor denotes those repetitive bodily behaviors of the human as biological animal that bind to nature: getting up and going to bed; the finding or growing, preparing, and consuming of food; washing and tending to body and clothing. For the *animal laborans* it is the species life that predominates and individuality barely exists. The species requires and only continues through labor, which takes form in cyclical patterns that echo those of nature: day and night; eating and eliminating; spring, summer, fall, and winter.

Second, work fabricates with the hands material things not found as such in nature, things that exhibit a measure of non-natural individuality and durability. Members of the species begin to manifest individuality in the mirror of the things they make: *homo faber* becomes a potter, an iron worker, a carpenter, a maker of weapons or household furnishings, a jeweler or painter. The durables of tools and buildings construct a world within which individuals are born and die and in the process pass from one generation to the next a more-than-biological culture: language, customs, traditions. Yet culture remains influenced or conditioned by the materials with which *homo faber* comes in contact. Desert peoples differ from mountain peoples; farmers from urban dwellers. As Arendt puts it, humans

> are conditioned beings because everything they come in contact with turns immediately into a condition of their existence. The world in which [humans live] consists of things produced by human activities; but the things that owe their existence exclusively to [humans] nevertheless constantly condition their human makers. In addition to the conditions under which life is given … and partly out of them, [humans] constantly create their own, self-made conditions, which, their human origin and their variability notwithstanding, possess the same conditioning power as natural things.... The impact of the world's reality upon human existence is felt and received as a conditioning force. The objectivity of the world—its object- or thing-character—and the human condition supplement each other; because human existence is conditioned existence, it would be impossible without things, and things would be a heap of unrelated articles, a non-world, if they were not the conditioners of human existence. (Arendt 1958, p. 9)

Latour may speak of artifacts as "society made durable" and of the impossibility of artifacts that do not incorporate social relations, but four decades earlier Arendt had already and with much less rhetorical excess described the emergence of durability and the mutual conditioning between humans and their world. The Churchill principle of Borgmann is a clear and sober reprise of Arendt's phenomenology of *homo faber* 50 years on, in a way that has the potential to nourish and inform a broader political discourse.

For within the world of artifice there emerges a third human activity, action, constituted by discourse among humans that is not subordinate to laborious cooperation in the family or apprenticeships of making in the workplace. Among humans who live in common beyond the ties of kinship there emerges *homo politicus* who, through a new kind of making, the making of laws, establishes an impersonal web of human relationships with its own distinctive durability.

> Labor assures … the life of the species. Work as its product, the human artifact, bestow a measure of permanence and durability upon the futility of mortal life and the fleeting character of human time. Action, in so far as it engages in founding and preserving political bodies, creates the condition for remembrance, that is, for history. (Arendt 1958, p. 8)

Action can become heroic and remembered not just through the writing of laws that establish a web of human relationships but also through contesting with—that is, acting into—a rival web. Indeed, rivalry in word and deed is coeval with the emergence of political action. And in a world in which action has through technology become "action into nature" (Cooper 1991; cf. Arendt 1958, pp. 231 and 324) there is an ever intensifying moral need for political discourse that can incorporate this new dimension of human affairs.

To repeat: Action gives birth to new things in human affairs—new things that have consequences, intended and unintended. We struggle to appreciate these consequences and be more than somnambulant makers of new worlds. What Winner argued in the 1980s echoes Arendt from the 1950s. As agents in the new world we must learn to reflect on the agencies we deploy through our actions. Only insofar as the discourse on agency and artifacts can contribute to this political demand will such academic efforts bear more than academic fruit.

Finally, in her distinctions between labor, work, and action Arendt invites us to recognize, considerably before Collins and Kusch, that agency takes many forms. One such form, action—insofar as it involves both deed and word—throws into relief the thinness of much contemporary discourse on agency and artifacts. Action does more that cause changes in the physical world; it reveals the uniqueness of actors who accompany deeds with performative words.

> Without the accompaniment of speech … action would not only lose its revelatory character, but … it would lose its subject …; not acting men but performing robots would achieve what, humanly speaking, would remain incomprehensive. Speechless action would no longer be action because there would no longer be an actor…. (Arendt 1958, p. 178)

Through her phenomenology of action Arendt cautions us to be careful, when reflecting on agency and artifacts, not to elevate artifice over speech by reducing speech itself to artifice.

## References


Adam, A. (2005). Delegating and distributing morality: Can we inscribe privacy protection in a machine? *Ethics and Information Technology, 7*, 233–242.

Akrich, M. (1992). The description of technical objects. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 205–224). Cambridge, MA: MIT Press.

Akrich, M., & Latour, B. (1992). A summary of a convenient vocabulary for the semiotics of human and nonhuman assemblies. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 259–264). Cambridge, MA: MIT Press.

Alexander, J. K. (2008). *The mantra of efficiency: From waterwheel to social control*. Baltimore: Johns Hopkins University Press.

Allenby, B., & Sarewitz, D. (2011). *The techno-human condition*. Cambridge, MA: MIT Pres.

Anders, G., & Eartherly, C. (1961). *Burning conscience*. New York: Monthly Review Press.

Anscombe, G. E. M. (1957). *Intention*. Oxford: Blackwell.

Arendt, H. (1958). *The human condition*. Chicago: University of Chicago Press.

Arthur, W. B. (2009). *The nature of technology: What it is and how it evolves*. New York: Free Press.

Averill, M. (2005). Unintended consequences. In C. Mitcham (Ed.), *Encyclopedia of science, technology, and ethics* (Vol. 4, pp. 1995–1999). Detroit: Macmillan Reference.

Bijker, W., & Law, J. (Eds.). (1992). *Shaping technology/building society: Studies in sociotechnical change*. Cambridge, MA: MIT Press.

Bimber, B. (1994). The three faces of technological determinism. In M. Roe Smith & L. Marx (Eds.), *Does technology drive history?* (pp. 79–100). Cambridge, MA: MIT Press.

Borgmann, A. (1984). *Technology and the character of contemporary life: A philosophical inquiry*. Chicago: University of Chicago Press.

Borgmann, A. (2005). *Real American ethics: Taking responsibility for our country*. Chicago: University of Chicago Press.

Brand, S. (1994). *How buildings learn: What happens after they're built*. New York: Viking.

Bratman, M. E. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Bratman, M. E. (2007). *Structures of agency: Essays*. New York: Oxford University Press.

Brown, M. B. (2009). *Science in democracy: Expertise, institutions, and representation*. Cambridge, MA: MIT Press.

Collins, H., & Kusch, M. (1998). *The shape of actions: What machines and humans can do*. Cambridge, MA: MIT Press.

Collins, H., & Yearley, S. (1992). Epistemological chicken. In A. Pickering (Ed.), *Science as practice and culture* (pp. 301–326). Chicago: University of Chicago Press.

Cooper, B. (1991). *Action into nature: An essay on the meaning of technology*. Notre Dame, IN: University of Notre Dame Press.

Davidson, D. (1980). Intending. In D. Davidson (Ed.), *Essays on actions and events* (pp. 83–102). New York: Oxford University Press.

De Certeau, M. (1980). *L'invention du quotidian*. Vol. 1, *Arts de faire*. Paris: Gallimard. English version: *The practice of everyday life* (trans: Rendall, S.). Berkeley, CA: University of California Press, 1984.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dreyfus, H. L. (1972). *What computer's can't do: A critique of artificial reason*. New York: Harper & Row.

Ellul, J. (1954). *La Technique ou l'enjeu du siècle.* Paris: A. Colin. English version: *The technological society* (trans: Wilkinson, J.). New York: Knopf, 1964.

Etzioni, A. (1968). 'Shortcuts' to social change? *The Public Interest, 12*(Summer), 40–51.

Etzioni, A., & Remp, R. (1973). *Technological shortcuts to social change*. New York: Russell Sage.

Feenberg, A. (1999). *Questioning technology*. New York: Routledge.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machine, 14*, 349–379.
Harbers, H. (Ed.). (2005). *Inside the politics of technology: Agency and normativity in the co-production of technology and society*. Amsterdam: Amsterdam University Press.
Hughes, T. P. (2004). *Human-built world: How to think about technology and culture*. Chicago: University of Chicago Press.
Illich, I. (1973). *Tools for conviviality*. New York: Pantheon.
Jerónimo, H., Garcia, J. L., & Mitcham, C. (2012). *Jacques Ellul and the Technological Society in the 21st century*. Dordrecht: Springer.
Johnson, D. G., & Powers, T. M. (2005). Ethics and technology: A program for future research. In C. Mitcham (Ed.), *Encyclopedia of science, technology, and ethics* (Vol. 1, pp. xxvii–xxxv). Detroit: Macmillan Reference.
Kelly, K. (2010). *What technology wants*. New York: Viking.
Latour, B. (1984). *Les microbes: guerre et paix suivi de irréductions*. Paris edition: Métailié, A.M. English version: *The pasteurization of France* (trans: Sheridan, A. & Law, J.). Cambridge, MA: Harvard University Press, 1988.
Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge, MA: Harvard University Press.
Latour, B. (1991). Technology is society made durable. In J. Law (Ed.), *A sociology of monsters: Essays on power, technology and domination* (pp. 103–131). London: Routledge.
Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 225–258). Cambridge, MA: MIT Press.
Latour, B. (1999a). *Pandora's hope: Essays on the reality of science studies*. Cambridge, MA: Harvard University Press.
Latour, B. (1999b). *Politiques de la nature*. Paris: Découverte. English version: *Politics of nature: How to bring the sciences into democracy* (trans: Porter, C.). Cambridge, MA: Harvard University Press, 2004.
Law, J. (Ed.). (1991). *A sociology of monsters: Essays on power, technology and domination*. London: Routledge.
Malle, B. F., Moses, L. J., & Baldwin, D. A. (Eds.). (2001). *Intentions and intentionality: Foundations of social cognition*. Cambridge, MA: MIT Press.
Mazlish, B. (1993). *The fourth discontinuity: The co-evolution of humans and machines*. Cambridge, MA: MIT Press.
Mitcham, C., & Waelbers, K. (2010). Technology and ethics: An overview. In J. K. B. Olsen, S. A. Pedersen, & V. F. Hendricks (Eds.), *A companion to the philosophy of technology* (pp. 367–383). Malden, MA: Wiley-Blackwell.
Mumford, L. (1964). Authoritarian and democratic technics. *Technology and Culture, 5*(1 Winter), 1–8.
Nader, R. (1965). *Unsafe at any speed: The designed-in dangers of the American automobile*. New York: Grossman.
Noble, D. (1977). *America by design: Science, technology, and the rise of corporate capitalism*. New York: Knopf.
Rosner, L. (Ed.). (2004). *The technological fix: How people use technology to create and solve problems*. New York: Routledge.
Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. New York: Cambridge University Press.
Searle, J. R. (1985). *Expression and meaning: Studies in the theory of speech acts*. New York: Cambridge University Press.
Shaw, J. (2006, June). Intention in ethics. *Canadian Journal of Philosophy, 36*(2), 187–223.
Simondon, G. (1958). *Du mode d'existence des objets techniques*. Paris: Méot. Second ed., Paris: Aubier, 1989.
Verbeek, P.-P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. University Park: Pennsylvania State University Press.

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. Chicago: University of Chicago Press.

Weinberg, A. (1966, December). Can technology replace social engineering? *Bulletin of the Atomic Scientists, 22*(10), 4–8.

Winner, L. (1980). Do artifacts have politics? *Daedalus, 109*(1), 121–136. Included with slight edits in Winner's *The whale and the reactor: A search for limits in an age of high technology* (pp. 19–39). Chicago: University of Chicago Press, 1986.

Wittfogel, K. (1957). *Oriental despotism: A comparative study of total power*. New Haven, CT: Yale University Press.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.

# Chapter 3
# Towards a Post-human Intra-actional Account of Sociomaterial Agency (and Morality)

**Lucas D. Introna**

**Abstract** In the history of ethical thought there has always been an intimate relationship between agency and questions of morality. But what does this mean for artefacts? It would not be too controversial to claim that the idea that artefacts have, or embody, some level of agency—even if it is very limited or derived in some way—has become generally accepted. However, there still seems to be wide disagreements as to what is meant by the agency of artefacts, how it is accounted for, and the subsequent moral implications of such agency. I will suggest that one's account of the agency of artefacts is fundamental to the subsequent discussion of the moral status and implications of artefacts, or technology more generally. In this contribution I will outline two different accounts of sociomaterial agency: (a) a human-centred inter-actional account (Johnson and VSD) and (b) a post-human intra-actional account (drawing on Latour, Barad and Heidegger). I will show that the post-human intra-actional account of sociomaterial agency posits the social and technical as ontologically inseparable from the start. Such a position has important implications for how one might understand sociomaterial agency and how one might deal with it. I will propose that the authors in the post-human approach all share what I call a 'co-constitutive' account of agency in which agency is not an attribute of the human or the technical as such but rather the outcome of intra-action. I will endeavour to illustrate the implications of such an account for our understanding of sociomaterial agency by considering the phenomenon of plagiarism detection. I will conclude by proposing disclosive ethics (in particular disclosive archaeology) as a possible way forward in dealing with the ethical and political implications of post-human intra-agencies.

L.D. Introna (✉)
Department of Organisation, Work and Technology, Lancaster University, Lancaster, UK
e-mail: l.introna@lancaster.ac.uk

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_3,

## 3.1 Introduction

Normative evaluations of artefacts and technologies are commonplace. For example, many people find weapons, nuclear technology and cloning—to name a few—morally or ethically problematic. Indeed, one often hears a particular technology or artefact being declared as 'good' or 'bad'. When making these evaluations people mostly have in mind the actual or anticipated consequences of the use of these technologies. They might suggest that technologies are just 'neutral' or 'passive' possibilities for doing things that only become morally significant when taken up by humans in line with their purposes (as represented in the slogan 'guns do not kill it is people that kill'—see Latour (1999)). This would suggest that it is the human purposes and actions that are morally problematic not the technology as such. Others would claim that technologies or artefacts, in their very design, allow (or prohibit) certain practices (and not others). As such they are morally significant from the start. In other words the moral question is already present in some way even before they are taken up in social practices. Irrespective of the direction one goes in locating the moral problem (i.e. human or technology), the claim that a particular artefact or technology is morally problematic presumes that it would therefore be desirable for us to *intervene* in some way or another to address this moral issue or problem. If this is true, then the next question would be to know how, where and when to intervene. In other words, such a possibility for intervention presumes that we can locate the distribution of morally significant agency in a given sociomaterial arrangement in such a way as to affect appropriate change. I would therefore claim that the question of sociomaterial *agency* is necessarily at the centre of any discussion of the moral status or implications of technology (as it is generally accepted to be at the centre of any discussion about moral issues in society more generally).

Now, most people would agree that artefacts or technology *does* things—a kettle boils water, a hammer drives in a nail, a computer sends an e-mail, etc. Thus, it would not be too controversial to claim that the idea that artefacts have, or embody, some level of agency—even if it is very limited or derived in some way—is generally accepted. What is disputed is the *nature and origin of that agency*. The difficulty with this inability to locate or account for sociomaterial agency in a straightforward manner is that we do not know how to go about addressing the normative and political issues that technologically mediated practices quite evidently raise. If the problem was simply that people tend to use technology in a normatively questionable way then we plainly have to govern the *use* of the technology more effectively (laws regulating access, training, etc.). If the problem, on the other hand, is the fact that the particular design of the technology allows for practices that are normatively questionable or undesirable then we need to regulate the *design* of technology more effectively (for example as suggested in value sensitive design). If however, sociomaterial agency is constituted in a more complex and subtle way, as I would suggest below (following Latour, Barad and Heidegger), then the issue of the politics and ethics of technology is itself constituted in more complex and subtle ways—i.e. it is

not open to simple intervention and correction (such as to regulate the use or to regulate the design). I would claim that without a satisfactory account of the constitutive nature of sociomaterial agency we will not be able to address adequately the normative and political implications of our increasingly technologically mediated sociality. More simply put: if we want to challenge, critique or change sociomaterial practices—normatively that is—then we need to know who (in terms of human and non-human actors) is doing what, when and how, i.e. we need to get a grip on the problem of the on-going constitution (or constitutive conditions) of sociomaterial agency. This is the aim of this contribution.

In what follows, I would like to explore, in a tentative way, the problem of sociomaterial agency and its moral implications. First, I will outline two different accounts of sociomaterial agency: a human-centred inter-actional account (Johnson and VSD) and a post-human intra-actional account (Latour, Barad and Heidegger). Second, I will use the post-human intra-actional approach to analyse the sociomaterial phenomenon of plagiarism detection. In doing this I will endeavour to show how the social and the technical is a co-constituted reality that is ontologically inseparable. Finally, I will propose the framework of disclosive ethics (in particular disclosive archaeology) as a way to deal with the ethical and political questions that our technologies raise.

## 3.2 Making Sense of Sociomaterial Agency (and Morality)

### *3.2.1 The Inter-actional Human-Centred Account of Sociomaterial Agency*

It seems clear that it is not feasible, given all the work that emerged from the STS tradition, and the philosophy of technology, to maintain a simple dualistic view of agency which claims that agency is located either in the human or in the artefact. It would be reasonable to say that there is a generally accepted view that agency is more distributed than such a dualistic view would suggest. Nevertheless, although there is this understanding that agency is more distributed, there is a group of scholars that believe it is important to locate (or believe we ought to locate) the original and most fundamental source of agency on the side of the human. In this regard I want to refer to two examples: a recent paper by Johnson (2006) on the moral agency of computers systems and the work on value sensitive design by Friedman et al. (2006) and Friedman and Nissenbaum (1996).

In her paper Johnson (2006) argues that computers are moral entities but not moral agents. Her argument is based on the notion that computers do not fulfil the basic criteria for moral agency as traditionally conceived, by for example Kant. In particular she suggests that the key to moral agency is the ‘intending act’ “because the intending to act arises from the agent’s freedom. Action is an exercise of freedom and freedom is what makes morality possible” (199). She continues to argue that

although computers do not exhibit 'intending acts'—which would make them moral agents—it does not follow that they do not embody intentionality. According to her computer systems have intentionality in that they embody the intentionality 'inserted' into them by the intentional acts of designers. She suggests that designers design systems to be poised to behave in certain ways. However, as she suggests, this is not the only intentionality at work. There is also the intentionality of the user. Thus, she concludes: "when computer systems behave, there is a triad of intentionality at work, the intentionality of the computer system designer, the intentionality of the system, and the intentionality of the user" (202) She proposes that all three of these intentionalities interact to shape the moral terrain that should become the focus of moral evaluation. Thus, according to her argument it would be a mistake—and misleading—to allocate moral agency to computers independently of human agency. Nevertheless, she proposes that it is ultimately human agency that should be the core focus of moral scrutiny: "when attention is focused on computer systems *as human-made*, the design of computer systems is more likely to come into the sights of moral scrutiny, and, most importantly, better designs are more likely to be created, designs that constitute a better world" (204, my emphasis). This is exactly what the value sensitive design (VSD) approach advocates.

Value sensitive design (Friedman et al. 2006; Flanagan et al. 2008) accepts the idea that technology embodies certain intentionality as proposed by Johnson. They claim that a particular design renders possible certain behaviours (in support of certain values) and not others. Proponents argue that the moral problem is that most designers work—often uncritically—with a limited set of values that represents the interests and values of a privileged subset of stakeholders—such as economy, efficiency, safety, and so forth. They argue it is possible to design technologies that embody and render possible a wider, more inclusive, set of behaviours and values. Like Johnson they accept an inter-actional human-centred view which suggests that: "values are viewed neither as inscribed into technology (an endogenous theory), nor as simply transmitted by social forces (an exogenous theory). Rather, the inter-actional 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. 2006, p. 361).

Central to the human-centred inter-actional account of sociomaterial agency is the view that all sociomaterial agency is originally human i.e. that it is humans doing things with or through technology. It is never technology doing things with or through humans as such. Furthermore, even if sociomaterial agency is not originally human in the full sense of the word we need to, or ought to, be able to trace it back to humans because we can only make humans morally responsible and accountable—i.e. they are the only fully fledged moral agents with the freedom to choose and to act originally. This need to locate moral responsibility in human agents is clearly an important requirement for us to organise and regulate society. However, I will suggest that although we might want to locate or allocate responsibility and accountability ultimately in this way for very good reasons we should not allow this moral (and pragmatic) requirement to unwittingly lead us into accepting a dualistic

account of sociomaterial agency. Or more fundamentally allow this requirement to lead us to accept an *ontology* in which we have to posit humans and technical objects as ontologically distinct entities (one intending and free, and the other not) which then interact to make sociomaterial entities possible. Besides the many philosophical controversies that such a view entails it must be said that the question of accounting for human agency as 'an exercise of freedom' is not unproblematic or uncontroversial.[1] How can we think of it otherwise?

### *3.2.2 The Intra-actional Post-humanist Account of Sociomaterial Agency*

The implied ontological dualism (and substantialism) in the inter-actional approach to sociomaterial agency has traditionally given rise to a number of now well articulated questions. For example, to what degree can the affordances/prohibitions of technology 'force' or make the user to do something? What about the intentions of the users? What about the variety of ways that users can interpret these technical affordances (Norman 1988)? What about unintended consequences never anticipated by the designers? More specifically, where are the normative significant questions 'located': is it in the artefact, in the user or in both? These are all very good questions. However, I would argue that these questions do not help us to get to grips with the complexity of sociomaterial agency as it happens in our everyday technology saturated lives. What is needed, I would argue, is a fundamentally different post-human account of sociomaterial agency. I will attempt to give such an account by drawing on the work of Latour, Barad and Heidegger in particular.

*Latour and the non-humans.* For Latour, as for Barad and Heidegger, any talk of humans and non-humans in ways that suggest that they are, separately, already what they are and then we 'add' them together to 'make' a sociomaterial world would

[1] Philosophers of action in the analytical tradition have asserted that an action, in some basic sense, is something an agent does that is *'intentional under some description'* (Donald Davidson 1980). They argue that there is a conceptual tie between genuine action, on the one hand, and intention, on the other. However tracking down the link between intention and action is not a simple matter at all—the large amount of work in action philosophy is testimony to this fact. In the continental tradition, especially in the work of Michel Foucault (1977) the original (or originating) subject is taken as deeply problematic. For Foucault subjects are the outcomes of discursive formations (constituted through prevailing power/knowledge regimes). Each regime of power/knowledge sustains a different type of subjectivity (i.e. the religious subject, the academic subject, the business subject, and so forth). If the original subject does not exist does it mean that particular 'subject centred' notion of agency does not make sense? Foucault would suggest not. To reject the autonomy (of the original subject) is not to reject agency. What is disputed is the necessary connection with an originating intention. Actions are intentional (under some description) but the intentionality does not originate in the subject and it transcends the subject in it being exercised. According to him there is often nobody (no specific actor) there to have 'invented' it as such (Foucault 1977). In social theory the relation between social structure and human agency has been a central and enduring problem as exemplified in the work of, for example, Anthony Giddens (1984).

simply be wrong. He claims: "There exists no relation whatsoever between the 'material' and 'the social world', because it is this very division which is a complete artefact…." (2005, p. 76). He further suggests that both humans and non-humans share a common history: "Humans and non-humans are engaged in a history that should render their separation impossible" (2003, p. 39). More than that, they do not merely share a common history; they are each other's common history: "A corporate body is what we and our artefacts have become. We are an object institution" (1999, p. 192). Very significantly to us he claims that in this institution that we are it is not a simple matter to allocate intentionality and properties this way or that way: "Purposeful action and intentionality may not be properties of objects, but they are also not properties of humans either. They are properties of institutions [collectives of humans and non-humans], apparatuses, or what Foucault called dispositifs" (1999, p. 192).

For Latour agency is distributed in such a way as to render it impossible to locate the sources of action in any precise way. He claims that an actor is not a source of action but rather the target of a vast array of entities that surround it. Action, he suggests, is "borrowed, distributed, suggested, influenced, dominated, betrayed and translated. If an actor is said to be an actor-network, it is first of all to underline *that it represents the major source of uncertainty about the origin of action*…" (2005, p. 46, emphasis added). This distributed, unoriginal, notion of agency should however not be seen as a 'weak' form of agency. Latour claims that when non-humans act as mediators they *make* other actors do things. He defines mediators as actors that *associate* with other actors in such a way that "they make others do unexpected things." (2005, p. 106). If agency is unoriginal, distributed and has power to "make others do things", as Latour suggests, then the issue of accounting for normative agency is indeed very important. In this regard Latour argues that if agency is distributed and not original to humans then so also is morality (i.e. those actions that are normatively significant):

> Morality is no more human than technology, in the sense that it would originate from an already constituted human who would be master of itself as well as of the universe… Morality and technology are ontological categories …and the human comes out of these modes, it is not at their origin. (2002, p. 254).

If Latour is right about the distributed and unoriginal agency of actors (or more specifically normatively relevant agency of actors) then one might conclude that it is ultimately impossible for us to deal with the ethical and political implications of electronically mediated social practices. One might conclude that 'following the actors' (as is often suggested by ANT scholars) will only continuously displace agency to somewhere else as we transverse the network of humans and non-humans—i.e. an infinite regress. I would suggest that this is where the work of Barad and Heidegger is important to help us account for sociomaterial agency in a way that may provide a way forward.

*Barad, phenomena and agential intra-action.* Barad's work is interesting as it emerges from the physical sciences, in particular her interpretation of the work of the physicist Niels Bohr and his attempt to find a convincing philosophical

framework to account for the seemingly contradictory results of quantum physics. For Barad (2003) the observer, her instruments of measurements and the objects observed are an ontologically inseparable unity, what she calls a phenomenon: "phenomena are the ontological inseparability of agentially intra-acting "components." That is, phenomena are *ontologically primitive relations*—relations without preexisting relata" (815, emphasis added). Phenomena are constitutive of reality, she argues. Barad (1996) proposes the notion of "intra-action" to deal with the fact that although phenomena are inseparable unities the two poles of the phenomenon (measuring apparatus and the object) do not exist as such apart from their ongoing intra-action. In other words there are not entities, which then interact. Rather, the entities are the performative outcome of the nexus of intra-acting relations—that is to say, these intra-acting relations are ontologically constitutive. In sociomaterial terms I take this to mean that the user/designer and the technological artefact or system is a phenomenon in which the social and the technical do not exist as such apart from their intra-action. In the nexus of intra-activity the phenomena are (re) produced: "phenomena are the place where matter and meaning meet" (Barad 1996, p. 185). Boundaries, between the social and the technical, are enacted and shaped through practices in intra-action, along with the phenomena. She suggests that "It is through specific agential intra-actions that the boundaries and properties of the 'components' of phenomena become determinate and that particular embodied concepts become meaningful. A specific intra-action (involving a specific material configuration of the 'apparatus of observation') enacts an *agential cut* … effecting a separation between 'subject' and 'object'" (Barad 2003, p. 815; italics in original). For our purposes I would rephrase this to mean that it is in specific agential intra-actions between users (and designers) and materiality that the boundaries and properties of the social and the technical becomes constituted as an ongoing intra-actional performativity (Butler 1993). Barad (2003) summarises her approach as follows:

> In summary, the universe is agential intra-activity in its becoming. The primary ontological units are not "things" but phenomena—dynamic topological reconfigurings/entanglements/relationalities/(re)articulations. And the primary semantic units are not "words" but material-discursive practices through which boundaries are constituted. This dynamism is agency. *Agency is not an attribute but the ongoing reconfigurings of the world*. (p. 818, emphasis added)

But what does this mean for responsibility? During intra-action, "marks are left on bodies. Objectivity means being accountable to marks on bodies" (Barad 2003, p. 824). For Barad the locus of responsibility is "a prosthetically embodied, performatively constituted agency" (Rouse 2004, p. 155) in which "we are responsible for the world in which we live not because it is an arbitrary construction of our choosing, but because agential reality is sedimented out of particular practices that we have a role in shaping" (Barad as quoted in Rouse 2004, p. 155). As Rouse (2004) suggests agency does not have to be an 'all-or-nothing' affair for us to take it seriously. Indeed, precisely because it is not an all-or-nothing affair do we need to subject the multiplicity of intra-actions, in concrete and specific practices of use and design, to meticulous analysis and scrutiny.

*Heidegger and sociomaterial being-in-the-world.* In *Being and Time* Heidegger argues that we humans (which he calls Dasein) exist in an ongoing structural openness 'with' the world in which we and the world are always already a unity, a being-in-the-world (Heidegger 1962, p. 297). We human beings (Dasein) are this unity or rather we have this unity as our ongoing way of being. Whenever we find ourselves or take note of ourselves, we find ourselves already in the world engaged in ongoing everyday activity in which things already and immediately show up as familiar 'possibilities for' this or that practical intention—never as mere objects that are just there. One could say their affordances are already immediately apparent to us. Indeed it is this prior apparentness that already makes them stand out as this or that particular thing in the first instance. Its location, arrangement, and all the implied references to a whole array of other things within the horizon of action (the already there referential whole) constitute it as 'obvious'—so we simply draw upon it in-order-to do what we want or need to do. However, when we take up these tools, as tools, we do not take them up for their own sake; we take them up with an already present reference to our projects or our concerns. As beings that have 'projectedness' (being already future oriented) as our way of being we find ourselves already immersed in a nexus of concerns that constitute us as that which we are or want to become. Or rather we have as our way of being a prior immersion in a nexus of concerns. This is why Heidegger (1962) claims the way of being of *Dasein* is care (care as in 'mattering') (p. 236). We encounter things in the world as mattering (being significant) because we matter to ourselves as being or becoming such or such a particular being (father, teacher, etc.).

Thus, we do not simply bang on keys, we use the laptop to type, in-order-to write this text, to do e-mail, to surf the web, etc. Moreover, the writing of this text already refers to the possibility of a presentation. This presentation in its being already refers to an audience, which refers to an institution, which refers to future audiences, which refer to research, which refer to further possibilities, etc. These references ultimately refer back to the being that I am or am becoming to be, i.e. a very particular being in the world of 'being an academic'. Heidegger (1962, p. 118) calls this recursively defining and necessary nexus of projects, or for-the-sake-of relations, the involvement whole. The equipment whole (of thing intra-relations) and the involvement whole (of care intra-relations) co-constitute each other—i.e. they are each other's transcendental condition for being what they are—in Barad's terms they intra-act each other. They sustain each other's way of being as an ongoing horizon of meaning. Heidegger calls this horizon of meaning 'the world'. The meaning (or coming into being) of us and our tools (the social and the technical) can only be understood within this already mutually defining referential whole, the world itself. Thus, as beings-in-the-world, our tools and us always already *co-constitute each other's possibility for being agents*—not in some general sense but exactly that which we are in this or that particular world (of academia, business, and so forth). But this is not all. If it is true that we exist in a co-constitutive relation with technology (also in more general terms) then our technological world is also more than just this or that particular co-constitutive practice (my word-processor and the academic me).

In other words there is a sense in which what it means to be human—and what counts as the real world—emerges from this co-constitutive whole.

In his essay "The Question Concerning Technology" Heidegger (1977) claims that: "Technology is therefore no mere means. Technology is a way of revealing. If we give heed to this, then another whole realm for the essence of technology will open itself up to us. It is the realm of revealing, i.e., of truth" (p. 12). Thus, for Heidegger technology is—in it co-constitutive becoming—the very disclosure of being.[2] Or as Ihde (1991) expresses it: "Technology, in the deepest Heideggerian sense, is simultaneously material-existential and cultural. …. It is a way of seeing [or being] embodied in a particular form" (Ihde 1991, pp. 56–57). One might say that in its ongoing becoming technology reveals, in a very fundamental manner, 'a way of being' in the world. That is why Heidegger (1971) claims in his essay *The Thing* that "the thing things world" (p. 181). Indeed that is the only way one can make sense of his suggestion that the "jug is not a vessel because it was made; rather, the jug had to be made because it is [already] this holding vessel" (p. 168). What we see is a seemingly 'reversal' of intentionality. The designer/craftsman did not decide (intend) to make the jug. The possibility of a jug was already suggested (intended) by the ongoing worlding of the world. The world (or referential whole) in which the jug, as a holding vessel, emerges as necessary is prior to this or that entity 'jug'. Therefore, in making the entity 'jug' a world (a way of being), already present, is revealed. As such technology—or precisely the technological way of being—has as its being the revealing of a way of being (an originating intentionality) that is prior to this or that artefact.[3]

Let me summarise what I suggest is Heidegger's post-humanist account of sociomaterial agency—what I would like to describe as *co-constitutive agency* (or what Barad will describe as intra-action)—by taking the CCTV camera phenomena as an

[2] Central to Heidegger's ideas is his notion of the 'ontological difference'. The ontological difference is the difference between being and entities. What an entity is depends on meaning-conditions that make entities stand out as that which it is. These conditions make up the *being* of entities. As Heidegger suggests "the being of entities 'is' not itself an entity" (Heidegger 1927/1962, p. 6); the being of entities is rather the implied conditions of possibilities (or horizon) against which entities make sense at all. Thus, the being of technology is not itself an artefact or system but rather the condition of possibilities against which artefacts emerge as meaningful. As such the being of technology reveals or discloses worlds.

[3] It is therefore no surprise that for Heidegger the essence of modern technology is the way of being of modern humans—a way of conducting themselves towards the world—that sees the world as something to be ordered and shaped in line with our projects, intentions and desires—a 'will to power' that manifest itself a "will to technology". It is in this technological mood that problems show up as requiring technical solutions. The term 'mood' here is used in a collective sense, like the 'mood of the meeting' or the 'mood of our times'. He calls this technological mood 'enframing' (*Gestell* in German). For us, in the technological age the world is already 'framed' as a world available 'to be made', 'to be shaped' for our ongoing possibilities to express our existence, to be whatever we are, as business men, engineers, consultants, academics, teenagers, etc. In short: the need for modern technology makes sense because we already live in the technological age or mood where the world (and us as beings that are never 'out' of the world) are already framed in this way—as available resources for the ongoing challenging and ordering of the world by us, which is for him the essence of the 'modern' mood.

example. A CCTV camera mounted on a wall can *make* humans—that want to see at a distance (or not be seen at a distance)—do what they do—zoom in, take note of suspicious behaviour; or, cover their faces, follow other routes, etc.—not because there is a particular cause (or agency) *in* the artefact as such (or in the human as such) but because CCTV cameras appear in the world of police officers wanting to see at a distance (or humans wanting to avoid being a surveillance target) as *already necessary and meaningful in that world of legal enforcement.* If the possibility of surveying at a distance (or not becoming a surveillance target) does not *concern* you or me then the CCTV camera might merely be a decorative object on the wall. Thus, the CCTV camera will only show up or stand out as something potentially relevant and meaningful in a nexus of concerns (and equipmentality) where the possibility of seeing (or not being seen) 'at a distance' might be taken as a *necessary condition* to realise the concerns that constitute the 'who' (the identity) that such a CCTV camera assumes or already refers to (the police officer or the person on the street that does or does not want to be targeted). *The important point is that the necessary or constitutive relation is not empirical as such, it is ontological—it renders possible the being-in-the-world of all the actors involved (camera, officer, suspect, etc.).* It is the necessary ontological co-constitutive intra-relation between cameras, operators and targets that renders sociomaterial agency possible in the empirical world of everyday action—i.e. which *makes* the actors do the things they do. Artefacts do script our behaviour in our dealings with them, as Latour suggests, but this 'scripting' is rendered possible by a prior, but already present, ontological co-constitutive intra-relation. Without such an intra-relation there is no script, no camera, no policeman and no suspect.

The condition of possibility for agency of all the actors (what we call the *co-constitutive agency*) is the always and already present horizon of meaningful possibilities to be—that which they suppose themselves to be—in the world. That is, the already present *necessary conditions* for a being (a CCTV camera, an alert police officer, a surveillance target) to be that which they are already taken to be in the world where they have their being. In saying this we must be careful to note that the constitutive horizon of the CCTV camera constitutes a multiplicity of actors (and identities) in the world it operates 'as a CCTV camera'. For example it constitutes what it means to be a police officer, what it means to be a 'suspect', how an officer relates to a 'suspect', what the prevention of crime means, and so forth. Furthermore, in and through the co-constitutive horizon (of CCTV cameras, police officers and surveillance targets) a particular understanding of the world (of crime, crime prevention, safety, security, etc.) is rendered possible and revealed as such. Thus technology, when it functions as such, reveals, in a very fundamental manner, 'a way of being' in the world (see also Introna (2009) for a more detailed discussion of the implications of this claim for human and non-humans).

Now that we have done a brief review of the post-human intra-actional account of sociomaterial agency I would like to consider the phenomenon of plagiarism detection in the world of learning and teaching to demonstrate how such an account might inform our understanding of the ethico-political implications of sociomaterial agency.

## 3.3 Figuring Intra-actional Agency in the Plagiarism Detection Phenomenon

In order to make the ethico-political implications of phenomena visible we need to do some figuring 'out' of the intra-actions. I want to suggest that we need to make some agential cuts to expose some of the 'components' or agencies that intra-act to constitute the being-in-the-world of plagiarism detection phenomenon. I want to propose—although I do not have space to defend this proposal here as such—that the following figuration agencies might be appropriate:

(a) *Affordances/ prohibitions*—The material affordances and prohibitions that constitute the form, fit and function of the material artefact (the computer algorithm, the word processor, electronic text, etc.) as well as that which constrains and enables the sort of affordances that may be imagined and rendered possible legitimately.
(b) *(Cyborg) Identities*—The ways of being someone in particular (teacher, student, author, plagiarist, etc.) as well as that which constrains and enables the sort of identities that can be assumed legitimately.
(c) *(Cyborg) Practices*—The ways of doing something in particular (writing an essay, evaluating an essay, reusing material, etc.) as well as that which constrains and enables that which can be done legitimately.[4]
(d) *Discourses*—The ways of talking (or making claims) about something in particular (what learning, assessment and academic writing is supposed to be, what plagiarism is, etc.) as well as that which constrains and enables that which can be said legitimately.[5]

These intra-actional agencies are in an ongoing co-constitutive intra-relation with each other to engender the ongoing becoming of the plagiarism detection phenomenon. Let us try and draw some brief and preliminary outlines of this phenomenon using the agencies above to figure it.

### *3.3.1 'Cutting and Pasting' and the Reconstitution of Writing and Authorship*

The automation of the construction of texts through the word processor reconstituted the practice of writing as well as the question of authorship in fundamental ways. For example Heim (1999) argues that in handwriting one's thoughts had to be thought through before being committed to the page—in other words that there is thinking and then writing. In contrast, he argues, when writing on the screen writing loses its reflective craft-like nature. According to him words and ideas on the screen

[4] Here I am using Rouse's (2007) normative conception of practice.
[5] Here I follow Foucault (1972, 1994) and his notion of discourse and discursive formations.

become constituted as fragments that can be 'cut and pasted' in a more or less thoughtless manner—the electronic text becomes constituted as never being thought as such. In the composition of electronic texts, he proposes, the relation between writing and thinking is reversed, more specifically, that there is writing and then thinking. Such an argument suggests that the text manipulation affordances of word processors such as 'cutting and pasting' not only makes the manipulation of text possible but it also reconstitutes the very practice of writing itself.

Moreover, when writing in an electronic media we find that authors do not just cut and paste *within* documents they also cut and paste *between* documents. As more and more texts became electronically constructed the idea of writing 'from scratch' becomes less and less attractive. In electronically mediated writing practices authors increasingly cut and paste from previously written texts—thus, we see the emergence of the practice of 'reuse.' This reuse is specifically implemented as the cutting and pasting of text 'as is'—which is of course different to transcribing. For example consultants 'reuse' parts of client reports, academics reuse written arguments developed in previous papers, lawyers reuse standard formulations in contracts, students reuse parts of earlier assessments, and so forth. In a world where efficiency has become a legitimate way of thinking about work the notion of reuse is enormously attractive (even normatively compelling). As such we find that 'reuse' of text by 'cutting and pasting' from previous documents emerges as apparent and familiar. Indeed doing it from scratch might even be seen as being wasteful. Furthermore, one could argue that the obviousness of textual reuse makes sense in a world where the practice of 'reuse' has already become the constitutive basis for many other authoring practices. For example in software programming code reuse has become the dominant approach. The paradigm of object oriented programming is based on the notion that certain standardised code (standard routines for doing things), or 'objects' as they are known, should be made available in a central repository for reuse. A good programmer is able to use these standard routines or objects to build complex applications. My point is that the seemingly simple affordance of word processors to allow for 'cutting and pasting' has not only made text manipulation possible (as may have been intended by the designers) but has intra-acted to reconstituted the whole act of writing through the notion of reuse—especially in a world where reuse has already become a legitimate (even normatively required) practice of 'being efficent'. Thus, what we increasingly see—especially amongst our students—is a form of writing that one might call *patch-writing* (Howard 1993, 1995). In patch-writing texts are constructed by using (or reusing) preformed fragments that can be cut and pasted from elsewhere as the basis from which the text becomes constructed—a very different practice of writing through which, or from which, thinking emerges rather than the other way around, as suggested by Heim (1999).

With the advent of the Internet (enabled by the search capability of for example Google), and electronic publishing, the database of electronic texts available for reuse has exploded. In the context of the availability (now on our desktop) of this massive database of electronic texts many authors, it seems, are increasingly not only cutting and pasting from their own previously constructed texts but also from

texts constructed by other authors. In doing this not only the practice of writing has become reconstituted but also the meaning of *what it means to be an author*. Such practice of using other author's texts seems quite legitimate in a world of efficiency where reuse and outsourcing (ghost-writers, speechwriters, etc.) is increasingly common (as has been in oral societies where stories were commonly owned and the notion of original authorship did not exist).[6] Furthermore, it seems that the question of reuse and outsourcing of textual fragments also makes sense to students in the context where the understanding of what education is (or supposed to be) has shifted with the increasing commercialisation and commoditisation of education (Saltmarsh 2004, 2005; Vojak 2006). Indeed, it is possible to see why students might think that if you pay for your courses why can you not also outsource the writing of your assessment—especially if you also have to hold down a part-time job to pay for your education (which turns out not to be 'part' time at all). Nevertheless, this reconstitution of the meaning of writing, authorship and education now emerges—especially in the university context—as the phenomenon of plagiarism—or more precisely the ethics and politics of *plagiarism*.

### *3.3.2 The Emergence of the Phenomenon of Plagiarism*

In many subjects assessment of the student's knowledge of the subject is understood as the ability to create an original text that reflects the student's own understanding of the ideas in the form of the academic essay. But what if these texts are increasingly the outcome of a reconstituted practice of patch-writing? What is the student that constructs such a text? What is it that they think they are doing? Are they authors or plagiarists? How is plagiarism understood in this intra-action of agencies?

The Oxford English Dictionary Online (OED Online) defines plagiarism as *"the wrongful appropriation or purloining, and publication as one's own, of the ideas, or the expression of the ideas (literary, artistic, musical, mechanical, etc.) of another.*" However, if we go back a bit further to Samuel Johnson's Dictionary of 1755 he defines a 'plagiary' as *"a thief in literature; one who steals the thoughts or writings of another" and "the crime of literary theft."* (Lynch 2002). It seems that the important difference between these two definitions is the notion of *"the expression of the ideas"* that seems to have been added by the Oxford dictionary to the 1755 meaning. The emphasis on 'expression' of ideas emerged later in the eighteenth century (Hesse 2002) as a way to allocate rights to authors (where 'expressions' are protected but not ideas). It seems that there has been a shift in focus from 'thoughts or writings' (i.e. ideas and works) to the notion of the 'the expression of the ideas' (exact copies of text). The emergence of this understanding

[6] The relationship between originality, authorship and ownership is a complex cultural and legal history of the rise of intellectual property rights which cannot be covered here (see Hesse 2002; Bracha 2006).

of plagiarism is central to the constitution of the contemporary plagiarism detection phenomenon as we shall see. It must also be said that there is very limited consensus in practice amongst academics and teachers as to what constitutes plagiarism, as a study by Roig (2001) indicated.

### 3.3.3 *'Cutting and Pasting' and the Constitution of the Plagiarist*

Plagiarism has always been an issue for universities. As suggested above, academic writing, the ability to construct an argumentative essay in response to a question that reflects ones understanding of a subject, has been at the heart assessment in the humanities and the social sciences for many years. Traditionally it was expected that any plagiarism by students would be picked up by the teachers involved when they tutor students in the writing task and when they mark or grade the essays. However, decreasing staff/student ratios as well as the sheer number of resources available to students has made this extremely difficult to achieve. In practice, what we find is that teachers tend to suspect plagiarism when they notice a sudden change in style (or voice) in the text. This happens most often with non-native speakers that lack the linguistic ability to integrate 'cut and paste' fragments into their patch-writing practices. The increased reporting of cases of plagiarism in the press as well as the availability of essay for sale on the web has created a situation of panic in which plagiarism detection systems (PDS) emerged as an obvious solution for universities (Lathrop 2000).

The market leader, *Turnitin*, claims that their system is used be 5,000 institutions in 80 countries worldwide (covering 12 million students and educators) and that 50,000 papers get submitted to their system every day. They also claim that their crawler 'Turnitinbot' has downloaded over 9.5 billion Internet pages to their detection database and that it updates itself at a rate of 60 million pages per day (Turnitin website). More recently academic publishers have also turned to Turnitin to help them protect themselves from publishing plagiarised material, which is obviously very damaging to their reputation (and profits one might add). Nevertheless, one of the most powerful arguments often put forward for adopting it (beyond resource constraints) is that it 'levels the playing field', indeed, that it is more fair than the hit and miss approach where individual teachers have to spot cases of plagiarism—it is what any fair teacher would do. The argument is made that teacher-based monitoring of plagiarism, as now constituted, tends to pick out weak students or non-native speakers because of the obvious shift in sophistication when a piece of plagiarised text is found embedded in an assessment document such as an essay or dissertation. But is it levelling the playing field or does it rather reconstitute a playing field that is even more uneven? I would argue that it is the latter. Moreover, that this is a much more serious issue since many of the important co-constitutive conditions (affordances) are now embedded in proprietary systems which are not open for scrutiny—an invisible micro-politics one might say. I would argue that in the phenomenon of

plagiarism detection Turnitin does not function merely as a technology to 'detect' plagiarists but rather as a phenomenon to co-constitute plagiarists (and what plagiarism is now seen to be) in morally significant ways. In the co-constitutive horizon of PDS the being-in-the-world of teaching, learning, writing, assessment and what it means to be a 'plagiarist' is constituted in such a way that it is difficult to track down and account for very significant "marks left on bodies" (in Barad's terminology).

If it is true that Turnitin covers almost all (if not all) of the web then anybody taking something from the web has an equal chance of being detected and that would most certainly be fair, a level playing field. However, what if Turnitin does not cover the entire web? In such a case the likelihood of somebody being detected would depend on whether they happen to take something from a place that Turnitin did (or did not cover). If Turnitin's claim that they cover 9.5 billion pages is true and the estimate that the web consists of 11.5 billion pages is correct (which would give them 83.6 % coverage) then one could argue that there is a relatively high probability that a student will be detected if they take something from the web. However these figures are misleading because a lot of the content that Turnitin needs to cover is in fact behind passwords (i.e. in the deep web), such as academic journals for example. In a small scale experiment we selected 103 fragments from a number of likely sources where students may take material from—in the publicly available as well as the deep web—and submitted it to Turnitin. Turnitin was only able to detect[7] 47 of these, a detection rate of 45.6 %. This experiment was repeated with a larger data set of 15,308 fragments. Of these Turnitin was only able to detect 48.4 %. If these results are to some extent generalizable (we are not claiming it to be at this stage) then a student taking something from the web has less than 50 % chance of being detected, which is quite low. My problem is not that some are caught and some get away, as it were. I am rather more concerned with the fact that Turnitin—in its increasingly pervasive status—has become the constitutive condition of what is seen as plagiarism and that most teachers are now beginning to think that a 'green light' from Turnitin means that a students has not cheated. In this constitutive horizon they often believe that those that are not detected by Turnitin are innocent and those that are detected are guilty. I would suggest that both of these assumptions are wrong or could be wrong. The first is partly wrong because of the partial coverage of Turnitin as indicated by our experiments. The second one might be wrong for more subtle and complex reasons, related to the operation of the *algorithm* and its interaction with patch-writing practices, which I now want to turn to.

One must first note that plagiarism detection software—contrary to what its name suggests—detects *copies not plagiarism*. How does it detect copies? A simple approach would be to compare a document character by character. However, this approach has a number of problems: (a) it is very time-consuming and resource intensive; (b) it is not sensitive to white spaces, formatting and sequencing changes;

[7] Detection here is defined as being outside of the 'green' zone in the originality report, i.e. having a correspondence of greater that 24 % with the texts in the Turnitin database. This percentage was determined by Turnitin themselves to compensate for incidental matches or false positives (which one would expect in a nine billion document database) and legitimate quotations.

and (c) it cannot detect part copies from multiple sources. To deal with these problems a number of algorithms have been developed. Unfortunately many of these (such as Turnitin) are now proprietary software and therefore not available for analysis and scrutiny. However, we have studied the logic of certain published algorithms, such as winnowing (Schleimer et al. 2003), as well as doing some preliminary experimental research of the way the Turnitin algorithm seems to behave. From these we are able to draw some important conclusions, which I will discuss below.

All detection algorithms operate on the basis of creating a digital 'fingerprint' of a document which it then uses to compare documents against each other. The fingerprint is a small and compact representation (based on statistical sampling) of the content of the document that can serve as a basis for determining correspondence between two documents (or parts of it). In simple terms the algorithm first removes all white spaces as well as formatting details from the document to create one long string of characters. This often results in a 70 % reduction of the size of the document. Further processing is done to make sure that sequences of consecutive groups of characters are retained and converted through a hash function[8] to produce unique numerical representations for each sequential group of characters. The algorithm then takes a statistical sample from this set of unique numerical strings (or hashes) in such a way as to ensure that it always covers a certain amount of consecutive characters (or words in our human terms) within a sampling window and stores this as the document's fingerprint.[9] A fingerprint can be as small as 0.54 % of the size of the original document.

From this very limited description of the algorithm it is clear that the detection algorithm is very dependent on certain characteristics of the copied text to remain intact for detection to be possible. In some cases a small amount of change in the right way (or place) will make a copy undetectable and in other cases a large amount of changing will still make it possible to detect. One of the key requirements for detection is that a *sufficiently long string of consecutive characters* from the original is retained in the copied version. The location, within the fragment, of the consecutive string is also important due to the sampling window. For example in experiments we did with Turnitin it became clear that if one would change one word in a sentence at the right place—often between the 7th and 14th word in the sentence—then Turnitin did not recognise it even if all the rest of the sentence remained exactly the same. Indeed we were also able to submit a fragment of 300 words where we changed approximately every 7th to 10th word and remain undetected. In contrast Turnitin detected a small fragment of 26 consecutive unchanged words. Given this behaviour of the algorithm it is possible for a student to incorporate large amounts of copied material by intentionally or unintentionally changing words in the right places in the text submitted and remain undetected—see also Heather (2010) for ways in which text can be rendered undetectable. Now my

[8] A more technical definition of hash function is "A hash function is a function that converts an input from a (typically) large domain [input values] into an output in a (typically) smaller range (the *hash value*, often a subset of the integers) (from http://en.wikipedia.org/wiki/Hash_function).

[9] Refer to Schleimer et al. (2003) for a more detailed discussion.

concern here is not to suggest ways that students might cheat. My concern is rather the way this behaviour of the algorithm might constitute an uneven playing field, especially for non-native speakers.

We know that non-native speakers learn to write by using fragments as 'patches' to imitate the vocabulary and structure of expressions as part of their transition to become competent in academic writing (Howard 1993, 1995; Shi 2004; Leki and Carson 1997). This is true not only for non-native speakers, it is also true for native-speaking academics when paraphrasing a difficult-to-understand text—even material within their own discipline. Roig (2001), in a fascinating study, provided college professors in psychology (all members of the American Psychological Society) with two different texts to paraphrase: the first was a difficult text from a peer-reviewed psychology journal article and the second was an easy-to-read text from an introduction-level psychology textbook. Twenty-six percent (26 %) of the professors appropriated text—strings of five words in length or more *without quotation marks*—from the original text, whereas only three percent (3 %) appropriated text from the piece that was easier to read. If psychology professors—and most probably native speaking students—feel the need to 'stay close' to the text when confronted with difficult material, we can see why, students who understand the importance of 'speaking' like the teachers and the people they read, do the same when it comes to doing their assessments. We also know that it is possible to use phrases and fragments from a text to say something completely different than that which the original author has said. Nevertheless, this is not my concern here; rather, my claim is that non-native speakers (and novices in a discipline) will tend to use larger fragments of consecutive words, for fear of losing the meaning, than native speakers and experts. Furthermore, native speakers (and novices) will tend to have the vocabulary and linguistic skills to make changes to the fragments without a loss of meaning—especially in the middle of sentences where it really matters from a detection point of view. Thus, it is my claim that non-native speakers (and novices) who appropriate fragments as part of their patch-writing practices will be disproportionately detected as opposed to native speakers—see also Pecorari (2003). This becomes even more problematic when administrators (rather than teachers) are used to identify cases of plagiarism using the Turnitin's 'originality report' traffic light system.[10]

### *3.3.4 PDS, Education and the Production of Intellectual Property*

There are many more intra-actions and agencies at stake in the phenomenon of plagiarism detection. For example the whole issue of intellectual property rights. When students' work becomes incorporated into Turnitin's database these essays

[10] Blue: less than 20 matching words; Green: 0–24 % matching text; Yellow: 25–49 % matching text; Orange: 50–74 % matching text; Red: 75–100 % matching text.

**Table 3.1** Summary of some of the intra-actional agencies that co-constitute the phenomenon of plagiarism detection

| Co-constitutive intra-actional agencies | Some examples |
|---|---|
| *Affordances/Prohibitions* | Word-processors, cutting and pasting function, electronic documents and databases, Google, Turnitin detection algorithm, virtual learning environments |
| *(Cyborg) Identities* | Being an author, concerned teachers, able students, producers of intellectual property, intentional/unintentional plagiarists, a good designer (Turnitin) |
| *(Cyborg) Practices* | Cutting and pasting, reusing, patch-writing, assessing learning, detecting cheaters, trading intellectual property |
| *Discourses* | Commoditisation of education, learning and teaching, cheating, fairness, authorship and originality, ownership and intellectual property rights |

partly enable Turnitin to perform its detection service (i.e. partly enables Turnitin to provide the service it charges for). In order to prevent legal problems universities ask students to sign agreements that their work can be submitted to Turnitin for purposes of plagiarism detection—i.e. sign away any property rights they might claim. Nevertheless, this very act of signing now constitutes the student as the *producer and owner of intellectual property*. Linked to this new identity is the increased value of 'original work' (now defined as that which the Turnitin system cannot detect). In this co-constitutive nexus students come to conceive of themselves as producing property (not doing an assessment) when they write an essay for a course assessment. Thus, in the context of the commodification of education (Vojak 2006) students quite naturally see themselves as producing intellectual property (now given extra value by Turnitin) to be sold in the open market. Hence, we now see students selling their essays and assessments on the internet (for example on e-bay). Moreover, in this constitutive context of assessments as 'property' and educational commodity markets we see the emergence of ghost writing services which can produce 'original work' that are guaranteed not to fall foul of the detection system.

Due to space limitations it is not possible to outline more of the co-constitutive agencies at work in the plagiarism detection phenomenon. Hopefully this brief sketch will at least indicate the potential of taking a different approach to sociomaterial agency. In Table 3.1 I summarise some of the co-constitutive intra-actional agencies at work in constituting the phenomenon of plagiarism detection in the educational context.

In summary: my suggestion is that the large-scale use of Turnitin may be creating a set of constitutive conditions or intra-actions in which some students are being constituted as 'plagiarists', and others not, in an unfair uneven playing field. Most importantly, and quite ironically, most of the teaching staff that use Turnitin are not

aware of this intra-action (and the intra-action of the plagiarism phenomenon more generally) and are contributing to it with the sincere intention to be fair. Moreover, a whole variety of practices, identities and discourses are being co-constituted through the ongoing intra-actional working out of sociomaterial agency in ways not anticipated or intended by any of the agents as such.

## 3.4 Intra-actional Agency and Disclosive Ethics

From our discussion of the plagiarism detection phenomenon above it is clear that the co-constitutive conditions (or intra-actions) that constitutes some students as 'plagiarists' (and others not) are *not* simply properties of software objects, but they are also not properties of the humans either. Indeed there is a fundamental co-constitutive agency at work in the nexus of intra-actional relationships. For example, we cannot say that the designers of Turnitin intended to discriminate against non-native speakers. The material agency of their code is but one element in the nexus of constitutive intra-actional relations. There is a multitude of other intentions and intra-actions at work that continues to render possible the ethico-political phenomenon or site in ways that transcend (even pervert) the intentions and affordances of any particular actant (in Latour's language). What we see in the intra-action is a reversal of intentionality. The teacher wanting (intending to be fair) adopts the affordances of PDS. The affordances of the PDS unfairly constitute some as plagiarists and others not. The outcome of the intra-action is that the agency of the teacher is one of arbitrariness or unfairness. Moreover, we cannot simply say that the software objects are neutral means and it is the people (teachers and students) which use them that are at fault, or that they simply use them in an inappropriate ways. Of course some of that might be true, however, the software objects do embody certain (im)possibilities, (dis)functions, affordances/prohibitions that condition the way they are taken up as part of ongoing social practices (in searching and detecting). Nevertheless, we cannot talk about affordances without already having to invoke all the other intra-actional agencies (identities, practices and discourses).

Does this mean we cannot 'locate' sociomaterial agency? We have suggested above that agency is not an all-or-nothing affair. We can make 'marks on bodies' visible. We can reveal the way in which these co-constitutive conditions intra-act to constitute some as plagiarists and others not (although our analysis above is incomplete). Nevertheless, through this brief analysis we believe we have shown that the morally significant location of agency is the phenomenon, a '*way of being in the world*' that acted as the ongoing co-constitutive horizon for the different actors (word processors, authors, plagiarists, teachers, students, etc.) to emerge in the way they did. I want to suggest that we need this type of disclosive analysis to help us make visible the nexus of co-constitutive intra-actions. I will refer to this as a *disclosive archaeology of the phenomenon* as part of a broader disclosive ethics approach (Introna 2007).

### 3.4.1 Disclosive Archaeology of Phenomena

Sociomaterial phenomena need to be subject to ongoing disclosive scrutiny through a process of disclosive archaeology as was briefly done with the plagiarism detection phenomenon above—and others such as search engines (Introna and Nissenbaum 2000), ATMs (Introna and Whittaker 2006), facial recognition systems (Introna and Wood 2004; Brey 2004) and virtual reality computer games (Brey 1999), to name but a few. When I use the term 'archaeology' here I am thinking of Foucault's work—i.e. the (transcendental) co-constitutive conditions that rendered a phenomenon possible. As he explains:

> … it is rather an enquiry whose aim is to rediscover on what basis knowledge and theory [sociomaterial agency in our case] became possible; within what space of order knowledge [sociomaterial agency] is constituted… Such an enterprise is not so much a history, in the traditional meaning of the word, as an "archaeology" (Foucault 1994, pp. xxi–xxii)

The purpose of disclosive archaeology is not to focus on material agency or human agency as such but rather to make visible the ongoing conditions of possibility, the way of being in the world, that render the co-constitution of agencies possible as part of the ongoing becoming of the phenomena. It must trace the contingent simultaneity of *affordances, identities, practices and discourses* to reveal the nexus that co-constitutes the ethico-political phenomenon or site of ongoing sociomaterial action—as was briefly sketched out above. But more than this it also needs to ask about the constitutive conditions that *constrains and enables* the sort of agencies (affordances, identities, practices and discourses) that can be imagined or emerge as legitimate in the nexus of co-constitutive intra-actions. In particular, what are the cultural historical conditions that enable and constrain the sort of affordances that is possible to conceive, the sort of identities that is possible to assume and the sort of practices that is seen as legitimate ways of acting? In our case example: how did it become possible for students to see education as a commodity? Why has academic writing and assessment become seen in the way that it did? Why did plagiarism and the need for plagiarism detection emerge? In other words, it is my claim that if we want to address the ethical and political questions that our technologies raise then we do not just need to address the affordances, identities, practices and discourses that constitute a particular sociomaterial phenomenon or site, we also need to ask about the constitutive conditions that enable and constrain the emergence of those particular agencies as legitimate in the first place.

### 3.4.2 Towards Intra-actional Responsibility

Having accounts of 'marks on bodies' is just one side of the equation; ultimately we need to act concretely in particular situations. In doing this we need to ensure that

we address all intra-actional agencies in its full simultaneity of intra-activity. For example we need to address simultaneously the:

- *Affordances/prohibitions*—We need to attempt to build values into the design of artefacts (as suggested by VSD) or materialise morality (as suggested by Achterhuis 1995; Latour 1991; Verbeek 2006). We also need to make artefacts more transparent so that the affordances and prohibitions of artefacts are more visible (Introna 2007; Winner 1980). We also need to build more engaging artefacts as suggested by Verbeek (2005) and Borgmann (1984). But more than this we also need to question the prevailing technological moods of our day. We must initiate, and participate, in the debates about the sort of technological futures we ought (or ought not) have.
- *(Cyborg) Identities*—When thinking about affordances we should also ask questions as to what sort of cyborgs we are becoming. We must participate in society more generally in developing technologically afforded notions of 'whole' identities rather than 'narrow' identities (such as gadget people, google generation, etc.). We must propose and show that technology can also afford the development of 'whole' identities within more mindful practices. In other words that all cyborg identities need not necessarily 'narrow.' But we also need to attend to the central question of what sort of cyborgs we want to become.
- *(Cyborg) Practices*—We need to understand the practices that are emerging around our technological affordances but we should also develop new technologically afforded (or cyborgian) practices that render possible our common human values. It is only in the nexus of practices of care (or mindfulness) that more mindful affordances can emerge as legitimate.
- *Discourses*—Most important of all is the development of new discourses that will enable and legitimate the sort of affordances, identities and practices that will intra-enact our common human values. Foucault was right when he said that discourses constitute 'subject positions' and naturalise them. I will add to this not just 'subject positions' but also, more specifically, technologically afforded identities and practices.

These suggestions are not complete, unproblematic or uncontroversial. Nevertheless, they seem to me to go some way in taking the ethics and politics of our increasingly sociomaterial existence seriously. More importantly, they attempt to acknowledge that agency is complex, distributed and not amenable to simple interventions (except in isolated and specifically constructed spaces/places). All socio-material interventions are mostly, if not always, more or less ontological in as much as they can reconstitute the agents (human and non-human) in many unexpected ways, as our archaeology of the plagiarism detection phenomena above revealed. The decision to take my car, or the bus, or my bicycle to work, constitutes me as a being that cares (or not) for the environment—and much much more. The question of morality in the constitutive nexus of socio-material phenomena cannot be resolved once and for all but needs to be worked out in the specifics of each constitutive nexus, again and again. This is indeed what gives ethics its urgency; there is indeed much work left to be done for us cyborgs.

## References


Achterhuis, H. (1995). De moralisering van de apparaten. *Socialisme en Democratie, 52*(1), 3–12.

Barad, K. (1996). Meeting the universe halfway: Realism and social constructivism without contradiction. In L. Nelson & J. Nelson (Eds.), *Feminism, science, and the philosophy of science* (pp. 161–194). Dordrecht: Reidel.

Barad, K. (2003). Posthumanist performativity: Toward an understanding of how matter comes to matter. *Signs, 28*(3), 801–831.

Borgman, A. (1984). *Technology and the character of contemporary life*. Chicago: Chicago University Press.

Bracha, O. (2006). *The ideology of authorship revisited*. Available at SSRN: http://ssrn.com/abstract=869446. Accessed April 2007.

Brey, P. (1999). The ethics of representation and action in virtual reality. *Ethics and Information Technology, 1*(1), 5–14.

Brey, P. (2004). Ethical aspects of face recognition systems in public places. *Journal of Information Communication and Ethics in Society, 2*(2), 97–109.

Butler, J. (1993). *Bodies that matter: On the discursive limits of "sex"*. New York: Routledge.

Davidson, D. (1980). *Essays on actions and events*. Oxford: Oxford University Press.

Flanagan, M., Howe, D., & Nissenbaum, H. (2008). Values in design: Theory and practice. In J. van den Hoven & J. Weckert (Eds.), *Moral philosophy and information technology* (pp. 322–353). Cambridge, UK: Cambridge University Press.

Foucault, M. (1972). *The archaeology of knowledge and the discourse on language* (A. M. Sheridan Smith, Trans.). New York: Pantheon Books.

Foucault, M. (1977). What is an author? (D. F. Bouchard & S. Simon, Trans.). In *Language, counter-memory, practice* (pp. 124–127). Ithaca/New York: Cornell University Press.

Foucault, M. (1994). *The order of things: An archaeology of the human sciences*. London: Routledge.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems, 14*(3), 330–347.

Friedman, B., Kahn, P. H., Jr., & Borning, A. (2006). Value sensitive design and information systems. In P. Zhang & D. Galletta (Eds.), *Human-computer interaction in management information systems: Foundations* (pp. 348–372). Armonk/London: M.E. Sharpe.

Giddens, A. (1984). *The constitution of society*. Berkeley: University of California Press.

Heather, J. (2010). Turnitoff: Identifying and fixing a hole in current plagiarism detection software. *Assessment & Evaluation in Higher Education, 35*(6), 647–660.

Heidegger, M. (1927/1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). New York: Harper & Row.

Heidegger, M. (1971). *Poetry, language and thought*. New York: Harper & Row.

Heidegger, M. (1977). *The question concerning technology and other essays*. New York: Harper Torchbooks.

Heim, M. (1999). *Electric language*. New York: Yale University Press.

Hesse, C. (2002). The rise of intellectual property, 700 B.C.—A.D. 2000: An idea in the balance. *Daedalus, 131*(2), 26–46.

Howard, R. M. (1993). A plagiarism pentimento. *Journal of Teaching Writing, 11*(2), 233–245.

Howard, R. M. (1995). Plagiarisms, authorships, and the academic death penalty. *College English, 57*(1), 788–805.

Ihde, D. (1991). *Instrumental realism: The interface between philosophy of science and philosophy of technology*. Bloomington: Indiana University Press.

Introna, L. D. (2007). Maintaining the reversibility of foldings: Making the ethics (politics) of information technology visible. *Ethics and Information Technology, 9*(1), 11–25.

Introna, L. D. (2009). Ethics and the speaking of things. *Theory Culture and Society, 26*(4), 398–419.

Introna, L. D., & Nissenbaum, H. (2000). The internet as a democratic medium: Why the politics of search engines matters. *The Information Society, 16*(3), 169–185.

Introna, L. D., & Whittaker, L. (2006). Power, cash and convenience: Translations in the political site of the ATM. *The Information Society, 22*(5), 325–340.
Introna, L. D., & Wood, D. (2004). Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance and Society, 2*(2/3), 177–198.
Johnson, D. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*(4), 195–204.
Lathrop, A. (2000). *Student cheating and plagiarism in the internet era: A wake-up call*. Englewood: Libraries Unlimited.
Latour, B. (1991). Technology is society made durable. In J. Law (Ed.), *A sociology of monsters: Essays on power, technology and domination* (pp. 103–131). London: Routledge.
Latour, B. (1999). *Pandora's hope. Essays on the reality of science studies*. Cambridge, MA/London: Harvard University Press.
Latour, B. (2002). Morality and technology: The end of the means. *Theory Culture and Society, 19*(5 & 6), 247–260.
Latour, B. (2003). The promise of constructivism. In D. Ihde & E. Selinger (Eds.), *Chasing technoscience: Matrix for materiality* (pp. 27–46). Bloomington/Indianapolis: Indiana University Press.
Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford: Oxford University Press.
Leki, I., & Carson, J. (1997). Completely different worlds: EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly, 31*(1), 39–69.
Lynch, J. (2002). The perfectly acceptable practice of literary theft: Plagiarism, copyright, and the eighteenth century. *Colonial Williamsburg: The Journal of the Colonial Williamsburg Foundation, 24*(4), 51–54.
Norman, D. A. (1988). *The design of everyday things*. New York: Basic Books.
Pecorari, D. (2003). Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing, 12*(4), 317–345.
Roig, M. (2001). Plagiarism and paraphrasing criteria of college and university professors. *Ethics and Behavior, 11*(3), 307–323.
Rouse, J. (2004). Barad's feminist naturalism. *Hypatia, 19*(1), 142–161.
Rouse, J. (2007). Social practices and normativity. *Philosophy of the Social Sciences, 37*(1), 46–56.
Saltmarsh, S. (2004). Graduating tactics: Theorizing plagiarism as consumptive practice. *Journal of Further and Higher Education, 28*(4), 445–454.
Saltmarsh, S. (2005). 'White pages' in the academy: Plagiarism, consumption and racist rationalities. *International Journal of Educational Integrity, 1*(1). http://www.ojs.unisa.edu.au/journals/index.php/IJEI/article/viewFile/17/65
Schleimer, S., Wilkerson, D., & Aiken, A. (2003, June). *Winnowing: Local algorithms for document fingerprinting*. In Proceedings of the ACM SIGMOD international conference on management of data (pp. 76–85).
Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication, 21*(2), 171–200.
Verbeek, P. P. (2005). *What things do – Philosophical reflections on technology, agency, and design*. University Park, PA: Pennsylvania State University Press.
Verbeek, P. P. (2006). Materializing morality – Design ethics and technological mediation. *Science Technology and Human Values, 31*(3), 361–380.
Vojak, C. (2006). What market culture teaches students about ethical behaviour. *Ethics and Education, 1*(2), 177–195.
Winner, L. (1980). Do artefacts have politics. *Daedalus, 109*, 121–136.

# Chapter 4
# Which Came First, the Doer or the Deed?

**F. Allan Hanson**

**Abstract** Two theories of action—methodological individualism and composite agency theory—are compared, together with their associated concepts of moral responsibility. They agree that deeds are done by doers, and that moral responsibility for a deed lies with its doer, but they differ on the definition of the doer. Methodological individualism holds that doers are limited to human individuals. Composite agency theory, noting that most deeds can be done only by humans working in concert with nonhumans (this is especially clear when computers are involved), defines a doer as whatever combination of human and nonhuman entities is necessary to accomplish a deed. Methodological individualism limits moral responsibility to human individuals while composite agency theory attributes it to the combination of humans and nonhumans that did the deed. Objections to this view of moral responsibility, and responses to them, are discussed. In the West, methodological individualism is shown to be rooted in humanistic modernity, while composite agency theory emerges from postmodernity. Nonwestern examples similar to both composite agency theory and methodological individualism are reviewed.

## 4.1 Introduction

On the day that I am writing this, the following appeared in a CNN Internet news report regarding the trial of actor Shelley Malil for attempted murder: "In his testimony, he stated that it was the knife that did it, and he stated this repeatedly." This defense—don't blame me, it was the knife—is not likely to gain much traction.

F.A. Hanson (✉)
Professor of Anthropology, University of Kansas, Lawrence, KS, USA
e-mail: hanson@ku.edu

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_4,

Virtually everyone would agree with victim Kendra Beebe, who said "In fact it wasn't the knife that stabbed me 23 times. It was Shelley Malil."[1] At the same time, he could not have stabbed her without the knife.

Eurotransplant is an automated system that rapidly generates priority lists of recipients for organ transplants on the basis of compatibility, age, waiting time, distance between donor and recipient, and balance among the several participating countries (Tufts 1996:1326).[2] Eurotransplant seeks to achieve "an optimal proportion between justice and efficiency—the medical ethical criteria," and it is generally thought that it realizes these objectives better than previous procedures that relied entirely on human evaluations (De Meester et al. 2000:333). Unlike Malil's knife, which was wielded by him with each thrust, once set in motion Eurotransplant can work largely by itself.

Situations such as these raise questions for both the theory of action and ethical theory. For the one, the question has to do with the nature of the agent, the doer of the deed. What is the place of the humans and nonhumans in the agencies that carry out the actions? For the other, the question is where to place responsibility for the actions? The ethical issue follows closely on that pertaining to agency, for virtually everyone agrees that (except when the agent is forced or is incompetent) responsibility for a deed falls upon its doer. But serious differences of opinion exist as to just who or what the doer is. These have been exacerbated in recent decades when, as a result of revolutionary developments in technology, it has become inescapably clear that most of what we do could not be done without the aid of computers and other nonhuman entities.

As for agency, two basic theories of social action may be distinguished, depending on the role that nonhuman entities are considered to have in it. "Methodological individualism" holds that only human beings are agents. The other theory of action, which I will call "composite agency",[3] holds that deeds can be done by any combination of human and nonhuman agents.

It is generally believed that for a deed to evoke moral responsibility it must fall under the ethical rules that cover behavior in a society, it has consequences for good or ill as defined in that society, and its doer must be aware of what is being done. Matters of morality, that is to say, are uniquely human, pertaining to human motivations and human evaluations of the quality of deeds. Moral responsibility poses no difficulty on the individualist account, for doers are limited to those human beings who possess the necessary awareness. In composite agency theory, on the other hand, the doer is usually not exclusively human. However, its nonhuman components lack the requisite awareness and thus the question arises as to their moral responsibility. Should Shelley Malil's knife, or the computers in the Eurotransplant system, share in

[1] http://www.cnn.com/2010/SHOWBIZ/Movies/12/17/california.actor.stabbing/index.html?hpt=P1&iref=NS1

[2] The participating countries are Austria, Belgium, Germany, Luxembourg, the Netherlands Slovenia, and Croatia.

[3] I now prefer this term over "extended agency," which I have used synonymously in earlier publications.

the moral responsibility for what they participate in doing? This brings a great deal of vexation to the issue of moral responsibility, and we will consider various attempts to deal with it as we go along.

## 4.2 Individualism

Methodological individualism has been deeply embedded in Western thought for centuries, with sufficient influence that it can be considered the standard or default position. Its premise is that all social behavior can be reduced to and explained in terms of the actions of human individuals (Jones 2000; Kincaid 1997; Udehn 2001). To its adherents this is so obvious as to be a matter of common sense. Consider, for example, the following statement by Anthony Flew (1995:61–62):

> All social collectivities are composed of individuals, and can act only through the actions of their components. Whatever is said about any mass movement, organized collectivity, or other supposed social whole, must at some stage be related and in some way reduced to discourse about the doings, beliefs, attitudes, and dispositions of its components. Who actually did and thought what; and what led them to act and to think, as in fact they did, and not otherwise?… All this, once it has been sharply stated, should appear obvious and altogether beyond dispute.

Flew here is mainly disputing the notion that human groups can, as groups, be considered as agents. Geoffrey Hodgson points out that many so-called methodological individualists in fact recognize that action should be explained in terms of both individuals and social relations among them, but he then goes on to argue that "methodological individualism" is scarcely the appropriate term for it (Hodgson 2007:220–221, 223).

Most important for our purposes, methodological individualism holds that computers, other machines, tools, and animals are not part of agency, but are objects that people encounter and manipulate in the course of their actions. This is the contention of numerous contributors to the debate, including Cohen (2000), Giere (2006), Himma (2009), and Matthias (2004). I would like to pay special attention to the 2009 essay "Artefacts without Agency," by Christian Illies and Anthonie Meijers. I select this because it explicitly discusses the important role of nonhuman entities (ranging from speed bumps to ultrasound) in action and in the moral responsibility for it. Illies and Meijers claim that their account is neither based on nor biased toward any specific theory of action or responsibility (439). I'm not sure what they mean by this, but between the individualist and composite agency theories of action discussed here, they unequivocally opt for the former. The word "agent" appears repeatedly in their essay, and every time it refers to human beings. Although they leave open the question of whether developments in artificial intelligence may lead to the attribution of agency to computers in the future, they state that for now "there are no compelling arguments to attribute agency to artefacts" (437).

The key concepts in Illies and Meijers' article are "action scheme" and "second-order responsibility." They use them to construct an argument that

limits agency and responsibility to humans while still recognizing an important role for technology and other artifacts. They prefer to speak not about particular actions, but about action schemes, which refer to the range of options available to people in various situations. Technology influences the choice of action by making certain options more or less attractive. They instance how a speed hump usually makes driving fast less attractive, thus influencing the person to drive more slowly (426–427). There is more to it than that, however, for they go on to explain how the physical circumstances, social expectations and personal motivations in play at the particular moment are important variables in determining the attractiveness of the technologically influenced options (427–431). In this way they can acknowledge the undeniable importance of nonhuman things in action while still limiting agency, the doer, to human beings who choose among and act upon the available options.

Moral responsibility enters the picture with the fact that some of the options in an action scheme, including those made possible by technology, are morally preferable to others (431). It would have been better for Shelley Malil to walk away from whatever confrontation he had with Kendra Beebe, or even to slap her, than to pick up a knife and stab her. Analysts say that using the computerized Eurotransplant system achieves a more just and equitable allocation of organs than operating without it. Illies and Meijers enrich the relation of artifacts and technology to moral responsibility with their notion of second-order responsibility (432–434). Some of the options for action available to people, and their attractiveness, result from the actions of other people. They use the example of an obstetrician making an ultrasound image of the fetus available to an expecting couple. The information provided by the ultrasound may change the attractiveness of options available to the couple regarding the fetus, leading them to act in a way they otherwise would not have done. The first-order moral responsibility for what the couple decides to do lies with them, but because the technology has changed their action scheme, a second-order responsibility is attributable to others: the physician, the engineers who designed the ultrasound test, and so on.

There may be difficulties in applying this model to our example of the Eurotransplant system. Second-order responsibility for its particular decisions rests, surely, with the manufacturers, programmers and operators of the computers. But where does first-order responsibility lie? Presumably humans can override any specific allocation of an organ that it makes. But that seems already to be at a remove from the ordinary decision-making process, which is an output of the system. The spirit of Illies and Meijers' account calls for some human being to make that decision, and therefore to have first-order responsibility for it. That does not seem to fit the facts.

Be that as it may, the important point is that for Illies and Meijers, moral responsibility always rests with human beings. Technology may have an important influence on what is done, but doers are always human, and the responsibility for what they do, be it first- or second-order, belongs exclusively to them. Hence the account by Illies and Meijers is individualistic, both in its theory of action and of moral responsibility.

## 4.3 A Modernist Frame

I have said that individualism is so deeply embedded in Western thought that it can be taken as the default position. The reason is that its social frame is humanistic modernity, which has (until very recently) been the dominant way of thinking in the West for at least 500 years. Humanism refers to the idea that the driving force in human affairs over that period has been the liberation and actualization of the autonomous human individual as the basic unit of society and the locus of meaning in life and in history. The movement was in full swing by the Renaissance, exemplified among numerous other achievements by Giovanni Pico della Mirandola's fifteenth century *Oration on the Dignity of Man.* It continued with the sixteenth century Reformation's idea that the individual has no need of mediation by the church, but stands in a personal relationship with God. What happened for faith in the Reformation was extended to knowledge in the Scientific Revolution of the seventeenth century, with its conviction that knowledge is to be acquired not from canonical texts, but from the observations and reason of the individual thinker. Humanism reached its apogee in the eighteenth century with the Enlightenment. One indication of the extreme individualism of this era is the notion that human beings originally lived in the state of nature as isolated individuals, who freely contracted with one another to form society. Another is the reduction of large scale economic developments to the behavior of individuals by theorists such as Adam Smith (Udehn 2001:7–10). Individualistic humanism retained its dominance until at least the mid-twentieth century, and it still remains the frame for liberal democracy, capitalism, and for the ideas of many social and philosophical thinkers.

The other element associated with modernity of importance to this analysis is the assumption of permanence and certainty. On the modernist view the furnishings of the world, all the way from atoms to complex life forms, are concrete objects that maintain their unique identities over time. One reason for their stability is that they are governed by laws of nature, themselves constant and immutable in their operation. The modernist perspective also holds that the forms of objects and the workings of natural laws can be known with certainty, primarily through the methods of science. The notion of the human individual as a thing that maintains its form and behaves with consistency through time is part of this overall worldview. Of course, no one holds that either the form or the behavior of an individual is immutable. Still, its continuity is readily recognizable as it does a variety of deeds over extended periods of time. Further, if human beings and other objects in nature change, the laws that govern those changes do not. Again, this is so deeply ingrained that it seems to be a matter of common sense. All of us are easily recognizable as the same persons today that we were yesterday, or 10 years ago. Joining this with the humanistic idea of liberation and autonomy, the individual emerges as an independent, free and autonomous being that persists in a more or less stable form, that engages in a series of deeds, and is responsible for what he or she does.

## 4.4 Composite Agency

Resistance to the individualism of humanistic modernity has been exerted at least since Rousseau who, while in many ways extremely individualistic, spoke of a general will that characterized society as a whole and that contrasted with the individual wills of its members. This general point of view was perpetuated by Comte, Durkheim, Kroeber, Leslie White and others, who held that a social or cultural level of organization, above the level of the individual, needs to be understood in its own terms. More recently, the conviction has been growing in social theory that the explanation of action should take into account the participation of those nonhuman elements that are necessary to the doing of the deed. This necessity has existed ever since our ancestors made the first hand axes in the Paleolithic, but the increasing reliance on electronic devices of all sorts in contemporary life has brought the matter into sharper focus. Because the deed could not be done without the participation of the nonhuman entities, they should be included with the human being(s) as parts of the agency, the doer of the deed. This point of view goes under names such as cyborg, actor-network theory, distributed cognition, and extended agency (Haraway 1991; Law 1999; Hutchins 1995; Verbeek 2009; Hanson 2004, 2007). For convenience, I will lump them all together under the name "composite agency theory."

One basis of composite agency theory, mentioned already, is the self-evident fact that humans do not and cannot act alone in order to accomplish what they do. It is possible to account for this within a methodological individualist theory of action. Illies and Meijers, as we have seen, acknowledge a role for technology and other nonhuman elements in terms of their impact on the attractiveness of the various options available to the human doer.[4] But composite agency theorists counter that it is insufficient to imagine that nonhuman entities are merely ancillary to the deed. If they are essential for the doing of the deed, then they should be considered to be part of the agency itself, that which does the deed. The growing prominence of computers in contemporary life makes it obvious to these thinkers that we must move beyond humanistic individualism. Bruno Latour writes: "To balance our accounts of society, we simply have to turn our exclusive attention away from humans and look also at nonhumans…. They knock at the door of sociology, requesting a place in the accounts of society as stubbornly as the human masses did in the nineteenth century. What our ancestors, the founders of sociology, did a century ago to house the human masses in the fabric of social theory, we should do now to find a place in a new social theory for the nonhuman masses that beg us for understanding" (Latour 1992:227). What is happening right here, right now, for example, can happen only because I am manipulating the keyboard of a computer. Myself, the computer hardware and the word processing software and the English language are necessary for the action to occur, so the doer in this case consists of all of us taken together.

A relatively early but highly provocative insight conducive to composite agency theory was articulated by Gregory Bateson. He said that the agent conducting any

[4] See Selinger et al. (2011) for a critique of Illies and Meijers's "action scheme" concept.

activity should be so defined as to *include* the lines of communication essential to that activity rather than cutting across them. He instances a blind man using a stick to walk down the street. The agency in this case should not be limited to the man but include all the essential communicating components: the man, the stick, and the street. Considered in this way, while it is clearly composed of concrete components, agency is also fluid because its components vary with the particular activity (Bateson 1972:459; see also Wood 1998; Hutchins 1995:291–292). Thus when that same blind man reads a book in Braille, the agent becomes the man and the raised markings on the page.

Hakken reinforces the point: "It is necessary to recast the objects of study, to no longer draw the boundary of the field's object at the human skin but treat humans and their technologies as unitary entities. A range of anthro-techno-science concepts (such as cyborgs and Creolized Objects) can help to do this" (Hakken 1999:224, see also Downey 1995:369). The fluidity of the doer can be brought out by identifying it not so much as a noun—an object or a collection of objects. It is more like a verb—an embodied activity, such as "a man reading a book in Braille." It is a notion consistent with physicist David Bohm's view of the world as informed by relativity and quantum theory, in which everything is an unbroken flow of movement and supposedly durable things such as observer and observed are only momentarily stabilized forms of movement that form wholes for a time and then flow apart to join in new configurations (Bohm 1980:xi, 47). It is a view radically at odds with the modernist concepts of fixed and stable objects discussed above.

The difference between individualism and composite agency theory which is most crucial to this analysis has to do with the definition of the doer. For individualism, as discussed above, the doer is the human individual, who precedes the deed and who remains essentially stable as it moves from one deed to another. Composite agency theory agrees with individualism that there can be no deed without a doer, but differs from it by claiming that there can also be no doer without a deed. If there were no deeds there would be no doers; in a world where nothing happens there are no agents. Thus the doer does not precede the deed, nor does it maintain a stable form through a series of different deeds. The doer is defined by the deed. Therefore the doer of any particular deed is that which does it, which is normally a composite agency. Such doers are not stable entities that do a series of different deeds. Different deeds define different doers. This is the fluidity of agency. The doer that reads a book in Braille is different from the doer that navigates along a street with a cane. These different concepts of agency—of the doer—account for the fundamental differences between individualism and composite agency, both as theories of action and, as we shall see in a moment, as moral theories.

The composite agency theorist's path from the definition of agency to moral responsibility passes through the concept of mind. Mind is manifested in actions, in deeds. We can say that the doer of those deeds has a mind only because it acts intelligently—with evaluation of the relevant circumstances, with foresight, prudence, sagacity and so on. Therefore, as with agency itself, "mind" should be understood more as a verb than a noun, a way of acting more than an object. This is the concept of mind advanced by the philosopher Gilbert Ryle when, critiquing Cartesian

dualism, he redefined mind from a nominative "ghost in the machine" to the verbal concept of intelligent performance (1949). Ten years earlier the anthropologist Leslie White made the same point when he argued that mind is "minding": not a thing but a way of behaving (1939). Clearly, that which "minds"—which acts intelligently—is the doer of the deed. After all, intelligent performances are just that: performances, or actions. Absent the performance there would be no intelligence, no mind.

Up to this point, methodological individualists and composite agency theorists could agree. The difference between them comes, again, with the definition of the doer. For methodological individualists the doer is the human individual. Composite agency theorists would acknowledge that some minding can be accomplished by human individuals acting alone, such as solo singing or daydreaming. But they would insist that most intelligent performances (playing a violin, writing a letter, shooting with a bow and arrow) require the participation of nonhuman objects. Because the participation of nonhuman objects is essential to the realization of the intelligent performance, it follows that the performer includes them as well as human participants. The performer or doer, that which minds, consists of the composite agency as a whole.

From here it is only a short step to moral responsibility. Again, deeds are central to the notion of moral responsibility, for there can be no responsibility for deeds which do not, or might not,[5] happen. Moral responsibility characterizes those deeds that have moral pertinence, that are judged to be good or bad. Everyone agrees that the moral responsibility for a deed lies with its doer. But the same difference between methodological individualism and composite agency theory as to the identity of the doer comes into play yet again. The methodological individualist, insisting that only humans are doers of deeds, limits responsibility to them. The composite agency theorist, claiming that the doers of deeds are often composite agencies consisting of both human and nonhuman components, attributes responsibility to those agencies as wholes. A methodological individualist might well insist that "guns don't kill people; people kill people." The composite agency theorist, noting the vastly increased likelihood that people with guns will kill people than people without them, would place moral responsibility on the agency consisting of the person and the gun. Similarly, moral responsibility for the organ allocations made by Eurotransplant rests with the system as a whole: its designers, operators, hardware and software.

Verbeek provides a starting point for this way of thinking by claiming that although nonhuman things do not have intentionality they do influence moral decisions (2006:121). So far, Illies and Meijers could (and do) say the same thing. But Verbeek goes on (in another publication) to say: "Ethics of technology is not a matter of juxtaposing the human activity of doing ethics and the non-human affordances of technologies that will affect human beings. It rather consists in linking the realms of the human and non-human, by taking technological mediations seriously

[5] I add "might not" to cover the responsibility for assuring that certain deeds or events do or do not happen in the future.

and actively 'styling' how they affect us" (2009:259). He points out how a technology such as ultrasound, which allows certain ethical questions to be raised about abortion, "immediately breaks the autonomy of the subject and also the purity of its will and its moral considerations." No longer can ethics be, as Kant would have it, exclusively a matter of reason apart from influence from the outside world, for now an outside world of technology is intimately tied with ethical decisions (247).

Numerous authors have elaborated on the proposition that artifacts such as computers are involved in moral responsibility. Deborah Johnson notes that action involving computer systems consists of three components working together: the user, the designer, and the computer itself. This triad corresponds to what is here called composite agency, and it has moral significance, for "computer systems cannot by themselves be moral agents, but they can be components of moral agency" (2006:203). This is especially clear in systems complex enough that "the distribution of tasks to computer systems integrates computer system behavior and human behavior in a way that makes it impossible to disaggregate in ascribing moral responsibility" (Johnson and Powers 2005:106). Van der Velden recommends that we should "understand ethical agency as emerging from particular sociomaterial configurations of people and artifacts" (2009:45). Hanson claims that there can be responsibility only for events that actually do or might happen, and because both the human and nonhuman components of extended agencies are necessary for the events to happen, they must share responsibility for them (2009:96).

And yet, the idea of attributing moral responsibility to anything other than human beings evokes resistance because it is so alien to conventional ways of thinking. This is largely due to two criteria that are traditionally deemed to be indispensable for moral responsibility: deserts (rewards and punishments) on the one hand, and mental qualities such as awareness, intention and foresight on the other. Both of these are imagined to be the exclusive province of human individuals. Let us consider them in turn.

Probably the first, knee-jerk reaction to the proposition that, for example, Kendra Beebe was stabbed by a composite agency consisting of Shelley Malil and a knife, is "That can't be so. We don't punish the knife!" Deserts are reserved for human individuals, and that is taken as evidence that they alone have moral responsibility. But the composite agency theorist could contend that this objection does not hold, for at least two reasons.[6]

First, it is simply not true that deserts are applied exclusively to human individuals. Training dogs includes praise and treats when the dog does as desired, as well as scolding and sometimes other punishments when the dog behaves improperly. In the Middle Ages the penalty for bestiality was hanging, and that included the animal as well as the human culprit. Moreover, the bells that were used to summon people to an uprising might be flogged or destroyed after the uprising was put down (Ihde 2006:273–274). When I was a boy my older brother and a friend slid down a steep hill on a toboggan. They hit a tree and my brother's friend suffered a severe

[6] See Hanson 2009:94–95.

broken leg. That evening his father came to our house with an axe, intent on destroying the offending toboggan.

These examples can be interpreted in two different ways. One is that deserts were applied (or not—my father talked the man out of chopping up our toboggan) because the animals, bells and toboggan were thought to be morally responsible for their part in the deeds. In that case, deserts can be taken as a necessary indicator of moral responsibility, but moral responsibility is not limited to humans. The other interpretation is that dogs, bells and toboggans are not considered to be morally responsible and deserts are applied to them for other reasons. Dogs, for example, might be rewarded to reinforce desired behavior. But then, because they can serve other purposes, reward and punishment are not necessarily indicators of moral responsibility. In either case, the proposition that makes a one-to-one link between deserts and human responsibility does not hold.

Perhaps more importantly, deserts are not a criterion for determining moral responsibility. It works the other way around: deserts are a response following decisions about moral responsibility; their application comes up only *after* that decision has already been made. Criminals are punished because they have done something morally unacceptable; they are not deemed to have done something morally unacceptable because they have been punished. Especially in cases potentially involving capital punishment, trials are conducted in two phases. The first is devoted to deciding whether the defendant is guilty of the crime. If the verdict is guilty, a second phase of the trial is convened to determine the punishment.

This is not to deny that deserts are applied for the most part to human beings. Certainly this is because they are apt to the human components of composite agencies in ways that they are not to nonhuman components. Deserts are meant to reward and encourage certain behaviors and to retaliate for and discourage others. This works only if the object of the deserts is aware that they are being applied, and why. Dogs may have this awareness to some degree, but tools, weapons, computers, and other objects don't know if they are being rewarded or punished. Thus doing so would be lost on them. This explains why deserts are applied to some beings and not to others, but it is no evidence for the proposition that deserts are a criterion for moral responsibility.

The other challenge to the proposition that composite agencies are morally responsible for what they do is that moral agents are commonly understood to know the difference between right and wrong, to have foresight and the capability of intending, deciding, feeling regret and satisfaction. The exceptions that prove the rule are those human beings who lack these qualities, such as small children and mental incompetents, who are not considered to be morally (or legally) responsible for what they do. Can this be squared with the composite agency theory's contention that the responsibility for a deed lies with its doer, which is often a composite agency?

The composite agency theorist's response to this has already been discussed. The qualities of intending, deciding and so on are all activities of mind. According to the argument above, activities of mind (or intelligent performances) are not limited to human individuals, but are often undertaken by composite agencies consisting of both human and nonhuman components. Such composite agencies may then be

considered to manifest the mental characteristics necessary for moral responsibility. To elaborate that argument somewhat, it is essential to recognize that matters of morality do not pertain to the natural world but are purely cultural. There was no morality before human beings had evolved to the point where some moral code was developed. Therefore the doer must include some rational human component if the deed is judged to have moral pertinence. This is why a bird of prey is not held morally responsible for eating lambs, nor a toilet that overflows and causes water damage because its shut-off valve malfunctioned. The mental criterion—the capacity to engage in intelligent performances—is also why small children and mental incompetents are not held to be morally responsible agents.

On the basis of this argument, the composite agency theorist can fully agree that morally responsible agents must have certain mental capacities and yet continue to hold that, because composite agencies may "mind"—may undertake intelligent performances—they qualify as having mental capacities and can therefore be held morally responsible for what they do.

To summarize our review, methodological individualism claims that composite agency theory is counterintuitive and that it introduces concepts and relationships that are unnecessary for understanding social action, mind, and moral responsibility. Composite agency theory retorts that individualism is outmoded and that the conditions of contemporary life require a different way of thinking that fully recognizes the interdependence of humans and nonhumans in virtually everything that happens. The modernist, humanist worldview that frames methodological individualism has already been discussed. We may now turn to the larger context or social frame that gives rise to composite agency theory.

## 4.5 A Postmodernist Frame

If the social frame for individualism is humanistic modernity, that for composite agency theory is the more recent postmodernity. In that frame, humanism—the notion that the fundamental unit in society and the driving force in history is the stable, autonomous human individual—is replaced by a more social, relational, and indeterminate view of things. In their review of the steps toward postmodernist social theory, Best and Kellner stress the structuralists' and poststructuralists' increasingly radical rejection of humanistic assumptions about the autonomous subject and unchanging human nature (Best and Kellner 1991:19–20, 27). David Gunkel agrees, noting how, after a long period of complete anthropocentrism in ethical theory, the challenge to humanism in the last three decades by structuralism and poststructuralism, together with other developments, has expanded ethics to include the treatment of animals and animal rights. The development of the field of machine ethics is a continuation of this expansion (Gunkel 2007:167–174).

Verbeek also signals the demise of humanism. He notes that the Enlightenment moved the source of morality from God to human beings. Now conditions of life have evolved to the point where we need to consider the morality of things

(2006:117). Deontological and consequentialist ethics represent humanist modernism (2009:245–247). "Both approaches take as their starting point a solitary human being that is either focused on the workings of its own subjective judgments, or on the objective consequences of its actions" (247). This separates human beings from and places them in opposition to objects. Nothing could be further from the reality of our contemporary world, where humans and objects collude and cannot be separated (245). Verbeek defends a posthumanism which "does move beyond humanism, but not beyond the human. It simply gives a central place to the idea that the human can only exist in its relations to the nonhuman. Not the *human* is declared obsolete by this form of posthumanism, but *humanism* as an all too human approach of what it means to be a human being. In order to cultivate humanity, we need to take seriously how also technologies help to cultivate us. Only by approaching the human as more-than-human it becomes possible to adequately give shape to the respect for humanity the humanist tradition has rightly been defending for so long" (261, his emphasis).

These are only a few of the scholars who hold that humanism and the primacy of the individual reflect an era in history that had a beginning and is now coming to an end (see Hanson 2004:468–470). Louis Althusser, for example, sees the individual as a contingent, constructed being. In his "Reply to John Lewis," Althusser argues that the engine of history for Marxism-Leninism is not "man" but the class struggle. Human nature is a variable product of particular forms of social relations; the individual as a transcendental agent struggling through history for freedom and independence is nothing more than the concoction of bourgeois ideology (Althusser 1976:46–54). Foucault puts a different twist on it by claiming that humanism's individual is the construction of Power/Knowledge in a particular historical era: "the very fact that we connect the different aspects of our being to make a coherent entity called the individual is the first and most significant thing that power does to us, making us feel vulnerable to judgment, as well as responsible for our behaviour, appearance and deeds, and the imaginary coherent and autonomous subjectivity they are supposed to reflect" (Mansfield 2000:110). Or, to quote Foucault himself in a poetic moment, the individual, born at a certain moment in history, may now be disappearing, "like a face, drawn in sand, at the edge of the sea" (Foucault 1970:386–387).

Edward Said explains Foucault's project in language that simultaneously illuminates the grip of the individual on previous social theory and identifies some of the sources of the growing disenchantment with it: "Classical European philosophy from Descartes to Kant had supposed that an objectively stable and sovereign ego (as in 'cogito ergo sum') was both the source and basis for all knowledge. Foucault's work not only disputes this but also shows how the subject is a construction laboriously put together over time, and one very liable to be a passing historical phenomenon replaced in the modern age by transhistorical impersonal forces, like the capital of Marx or the unconscious of Freud or the will of Nietzsche" (Said 2000:16; see also Kincaid 1997:2). Elsewhere Said cites, in addition to Foucault, Levi-Strauss, Barthes, and Lacan as also discerning the end of the subject (1985:292–293), and one may add Deleuze and Guattari (1983) as well.

Contemporary theory in a number of fields is shifting the unit of action and analysis from the stable, Cartesian individual to fluid information networks or composite agencies consisting of multiple human beings plus various nonhuman elements. Psychologist Kenneth Gergen suggests that "we may be entering a new era of self-conception. In this era the self is redefined as no longer an essence in itself, but relational" (Gergen 1991:146). Psychologists and other social scientists have also questioned the autonomy of the human individual via the concept of distributed or socially shared cognition (Resnick et al. 1991; Derry et al. 1998; Moore and Rocklin 1998). Thus Jean Lave analyzes learning as a social rather than an individual phenomenon (1991:64), while Edwin Hutchins painstakingly demonstrates how the computational process of navigating a ship can be fully understood only in terms of teams of individuals coordinating their several activities with each other and with various technological instruments (Hutchins 1991, 1995). As Lucy Suchman put it, "Agency—and associated accountabilities—reside neither in us… or in our artifacts, but in our intra-actions" (1998:12).

Yet another influential challenge to the centrality of the individual goes under the name of actor-network theory. As developed by students of science, technology and society, this theory attributes agency not to human individuals but to networks, finite in duration and variable in composition, defined according to the activity under analysis (Callon 1987:93, 1999:182–183; Law 1999:3–7; Latour 1987:84, 89, 1988). These actor-networks include, in addition to human beings, a wide variety of nonhuman components (Law 1991:10–11, 16–17; Star 1991:32–33). Similarly, David Gunkel holds that communication, which involves multiple individuals and is often mediated by electronic or other technological devices, is the province of recombinant cyborgs (Gunkel 2000:340). "Borg subjectivities…are not conceptualized as preexisting, selfsame, or self-determining individuals. Rather, they are relational subjects constructed and reconstructed based on the vicissitudes of the network…. Borg subjects float, suspended between points of objectivity, being constituted and reconstituted in different configurations in relation to the discursive arrangement of the occasion" (345).

## 4.6 Zooming Out

All these contributions share the notion that agency is multiple, relational, fluid and recombinant, assuming different configurations defined by the various activities that it undertakes. In this, composite agency theory is a product of the postmodernist emphasis on fluidity and indeterminacy, and it contrasts sharply with the modernist assumptions of stability and certainty that frame individualism. And yet, late twentieth century postmodernism and earlier humanistic modernism are not the only contexts that frame composite and individualist theories of action and morality. On the composite side, Burckhardt noted that social relationships held priority over individual autonomy in the medieval European concept of the self (Burckhardt 1956 (1860):100–101). Spinoza's view of God as immanent in nature rather than a

transcendent creator who rules over nature like a king over his kingdom collapses the distinction between doer and deed in a way similar to the present discussion of composite agency. "Thinking through Spinoza's treatments of individuality and of freedom shifts attention from concern with who did what, and to what end, to seeking a better understanding of what is done and what we are who do it. It shifts our attention to the circulation of images and affects embedded in social practices. The loci of responsibility shift from individuals to social practices and institutions" (Gatens and Lloyd 1999:72). Gatens and Lloyd build upon Spinoza's perspective to consider how the individual (they use the term "self") is the product of a complex conjunction of historical traditions (as an anthropologist, I would say "culture") and personal experience acquired through interaction with others. Because the self is so heavily defined by influences external to the individual, people often feel a sense of responsibility for things in which they are not directly involved, and which indeed may have happened before they were born. They instance a corporate responsibility recognized by contemporary Australians for the historical dispossession of land from the Aborigines (143–146).

Similarly, Nietzsche (often recognized as a precursor to postmodernism) embraced the notion that the doer is defined by the deed. In *The Genealogy of Morals* (First Essay, Section 13, 1956) Nietzsche demonstrates this with two examples. The first is that lightning is not two things, the lightning and the flash it produces. There is just one thing, for lightning *is* the flash (or the flash *is* the lightning). The second example concerns a bird of prey that eats a lamb. It is not, holds Nietzsche, that first there is a bird of prey which then, among other things, eats lambs. Instead, the deed defines the doer: eating lambs is what makes it a bird of prey.[7]

A number of non-Western cultures also understand agency in ways similar to the indeterminate, fluid view of composite agency theory. In Melanesia and aboriginal Australia the person is defined as much by position in a network of social relations as by individual traits (Strathern and Stewart 1998; Wagner 1991; Myers 1986). Confucius held that "persons are not perceived as superordinated individuals—as agents who stand independent of their actions—but are rather ongoing 'events' defined functionally by constitutive roles and relationships as they are performed within the context of their specific families and communities" (Ames and Rosemont 1998:20). As Huston Smith put it, the Confucian view is that "apart from human relationships there is no self. The self is the center of relationships. It is constructed through its interactions with others and is defined by the sum of its social roles…. Confucius saw the human self as a node, not an entity; it is a meeting place where lives converge" (Smith 1991:180).

[7] Nietzsche uses the example to make a point about ethics: if one distinguishes a being from what it does, one can then say it shouldn't do it and condemn it for an immoral action. In that way the weak manage to emasculate the strong. This may be so, but I suggest that Nietzsche's use of the bird of prey to prove it is misplaced because morality is grounded in rules for behavior and judgments about good and bad, right and wrong. These occur only in the realm of human culture and do not apply to purely natural events such as the behavior of birds of prey. Nevertheless, his definition of the doer in terms of the deed remains isomorphic with a composite agency theory of action as it has been presented here.

The same perspective is found in Zen Buddhism: "seeking after and grasping at a 'coherent self' that is non-existent from the outset only leads to a 'suffering.' The Buddhist idea of 'codependent arising' maintains that all things under the sun arise in a codependent relationship with each other. Nothing in the world exists in complete independence and isolation from others. There is no such a thing as a solid basis that exists autonomously" (Nishigaki 2006:240).

Yet another example of the relational self is found in the Inuit view of the relations between humans and animals. The Inuit believe that animals voluntarily present themselves to be hunted and killed. The hunter then has the obligation to perpetuate that gift by sharing the food with other people. Insofar as they do so, the cycle of gift-giving continues, but if humans do not share what they kill, the animals will withhold themselves and the hunt will fail (Gombay 2010:243). As Gombay interprets this, quoting Osteen's study of the gift, "By giving and receiving, then, we are linking ourselves to others. 'We cannot understand the gift if we persist in the idea that gifts are given and reciprocated by autonomous individuals… because in giving and receiving we expand the self, paradoxically, by firmly attaching it to social relations' (Osteen 2002:33)" (243).

This is not to say that variants of composite agency theory are ubiquitous outside the West. Native North America provides several striking examples of individualism. Walter Goldschmidt depicts the hunting-and-gathering Yurok and Hupa of Northwestern California as highly capitalistic, with individual ownership of property and an ethics focused squarely on the individual. He compares their sense of responsibility and their personal character structure to the European Protestant ethic: "Northwest Californian ethics placed the focus of moral responsibility upon the individual, a moral responsibility which internalized the command to industriousness, self-denial and personal aggrandizement; a moral demand which produced a pattern of individual guilt and the concept of sin" (Goldschmidt 1951:518).

The Central Algonkian Fox of the western Great Lakes region were more individualistic than any European humanist. They were extremely reticent about allowing anyone to have authority over anyone else (even fathers over their sons), and they rejected any kind of privileged social hierarchy (Miller 1955). "To early European observers the Fox individual appeared unusually haughty, self-contained, and quick to resent anything he perceived as limiting his right to independent action" (286).

Miller's account is particularly interesting because he describes the social frame or general cultural context for Fox individualism, particularly as found in concepts of the supernatural. Supernatural beings are natural phenomena of any sort: Skunk, Eagle, Fire, Corn, Swamp, and so on. There is no established hierarchy among them. They all draw upon a "generalized essence of supernatural power" called *Manitu*. But, in an interesting parallel to postmodern fluidity and indeterminacy, possession of this power is not permanent. The myths describe the struggles of these beings with each other, but the victors derive no lasting advantage because in subsequent conflicts the result may be different (279–280).

The relation between humans and the supernatural beings is similar. Upon reaching adolescence a Fox male undertakes a vision quest during which a supernatural being

comes to him and takes him under its protection. It is a reciprocal relationship, for supernatural beings crave tobacco but can get it only from the humans, who provide it to maintain the guardianship. But the supernatural guardian also owes something to its human protégé. If the man should fail in his enterprises he might well decide that his guardian is ineffective and abandon it for another (280–281). Such was the case with an incident in the effort to convert the Fox to Christianity.

> In 1671 warriors, undertaking a war expedition against the Sioux, painted the cross on their bodies and shields, put themselves under the protection of the cross-manitu, and gained a decisive victory over their enemies. They returned, proclaiming the white man's manitu. The following year, however, another expedition against the Sioux, under similar manitu protection, was disastrously defeated. In a rage, the warriors repudiated the white man's manitu, tore down the cross Allouez had erected, and refused to let the priest re-enter the village (281).

## 4.7 Conclusion

Returning to the West, the fulcrum in the debate between individualism and composite agency as theories of action and morality is the relation between the deed and the doer. Individualism's answer to the titular question of this essay is unequivocal: the doer comes before the deed. The doer is present both before and after the deed, and therefore exists separately from it. Composite agency's retort is that neither the doer nor the deed comes first because they are mutually dependent. There is no deed without a doer, and no doer without a deed. Individualism's separation of doer and deed is replaced by the notion that the deed defines the doer.

These different understandings of agency generate sharply different theories of action. Individualism is born of modernity. The humanistic character of modernity places the individual at the center of action, while the concept of the individual as a stable being that moves relatively unchanged through a series of activities is allied with modernist assumptions about the fixity or permanence of the furnishings of the universe, the natural laws that govern them, and the human possibility of attaining certain knowledge of them. With postmodernity assumptions about stability and certainty are replaced with a pervasive sense of flux, contingency and indeterminacy. Agency becomes as much a verb as a noun, a doing of something. From this perspective, that which is done has a determinative influence on that which does it.

The issue of moral responsibility seems more vexed than that of agency, especially on the composite side. But it is actually a secondary issue that flows directly from action theory. As stated at the beginning, both individualism and composite agency theory would affirm that there can be no responsibility for a deed that does not happen, and that responsibility for a deed lies with its doer. They allocate responsibility differently because they define the doer differently. If, as individualism would have it, the doer—the agent—is a human being and nothing more, then moral responsibility lies there and nowhere else. But if, as composite agency theory contends, the doer consists of all those elements necessary for the deed to occur,

then responsibility for it, both causal and moral, is shared by all of them. This fundamental difference governs the debate over whether moral responsibility must be limited to human beings or can include composite agencies consisting of both human and nonhuman elements.

**Acknowledgement** I am grateful to Rex Martin, Richard DeGeorge, Deborah Johnson, Evan Selinger and Louise Hanson for their penetrating and very helpful comments as this essay took its present form.

## References

Althusser, L. (1976). *Essays in self criticism*. London: NLB.
Ames, R., & Rosemont, H. (1998). *The analects of Confucius: A philosophical translation*. New York: Ballantine.
Bateson, G. (1972). *Steps to an ecology of mind*. New York: Ballantine.
Best, S., & Kellner, D. (1991). *Postmodern theory: Critical interrogations*. New York: Guildford Press.
Bohm, D. (1980). *Wholeness and the implicate order*. London: Routledge & Kegan Paul.
Burckhardt, J. (1956 (1860)). *The civilization of the renaissance in Italy*. New York: Modern Library.
Callon, M. (1987). Society in the making: The study of technology as a tool for. In W. E. Bijker, T. P. Hughes, & T. J. Pinch (Eds.), *The social construction of technological systems: New directions in the sociology and history of technology* (pp. 83–103). Cambridge: MIT Press.
Callon, M. (1999). Actor-network theory–The market test. In J. Law & J. Hassard (Eds.), *Actor network theory and after* (pp. 181–195). Oxford: Blackwell.
Cohen, R. A. (2000). Ethics and cybernetics: Levinasian reflections. *Ethics and Information Technology, 2*, 27–35.
De Meester, J., et al. (2000). In the queue for a cadaver donor kidney transplant: New rules and nephrology dialysis. *Transplantation, 15*, 333–338.
Deleuze, G., & Guattari, F. (1983). *Anti-Oedipus: Capitalism and schizophrenia*. Minneapolis: University of Minnesota Press.
Derry, S. J., DuRussel, L. A., & O'Donnell, A. M. (1998). Individual and distributed cognitions in interdisciplinary teamwork. *Educational Psychology Review, 10*, 25–56.
Downey, G. L. (1995). Human agency in CAD/CAM technology. In C. H. Gray (Ed.), *The cyborg handbook* (pp. 363–370). New York: Routledge.
Flew, A. (1995). *Thinking about social thinking*. Amherst: Prometheus Books.
Foucault, M. (1970). *The order of things: An archaeology of the human sciences*. New York: Vintage.
Gatens, M., & Lloyd, G. (1999). *Collective imaginings: Spinoza, past and present*. London: Routledge.
Gergen, K. J. (1991). *The saturated self: Dilemmas of identity in contemporary life*. New York: Basic Books.
Giere, R. N. (2006). *Scientific perspectivism*. Chicago: University of Chicago Press.
Goldschmidt, W. (1951). Ethics and the structure of society: An ethnological contribution to the sociology of knowledge. *American Anthropologist, 53*(4), 506–524.
Gombay, N. (2010). Community, obligation, and food: Lessons from the moral geography of Inuit. *Geografiska Annaler: Series B, Human Geography, 92*(3), 237–250.
Gunkel, D. J. (2000). We are Borg: Cyborgs and the subject of communication. *Communication Theory, 10*(3), 332–357.
Gunkel, D. J. (2007). Thinking otherwise: Ethics, technology and other subjects. *Ethics and Information Technology, 9*, 165–177.

Hakken, D. (1999). *Cyborgs@cyberspace? An ethnographer looks to the future*. New York: Routledge.
Hanson, F. A. (2004). The new superorganic. *Current Anthropology, 45*, 467–482.
Hanson, F. A. (2007). *The trouble with culture: How computers are calming the culture wars*. Albany: SUNY Press.
Hanson, F. A. (2009). Beyond the skin bag: On the moral responsibility of extended agencies. *Ethics and Information Technology, 11*, 91–99.
Haraway, D. J. (1991). *Simians, cyborgs, and women: The reinvention of nature*. London: Free Association Books.
Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*, 19–29.
Hodgson, G. M. (2007). Meanings of methodological individualism. *Journal of Economic Methodology, 14*, 211–226.
Hutchins, E. (1991). The social organization of distributed cognition. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 283–307). Washington, DC: American Psychological Association.
Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.
Ihde, D. (2006). Forty years in the wilderness. In E. Selinger (Ed.), *Postphenomenology: A critical companion to Ihde* (pp. 267–290). Albany: SUNY Press.
Illies, C., & Meijers, A. (2009). Artefacts without agency. *The Monist, 92*(3), 420–440.
Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*, 195–204.
Johnson, D. G., & Powers, T. M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and Information Technology, 7*, 99–107.
Jones, R. H. (2000). *Reductionism: Analysis and the fullness of reality*. Lewisburg: Bucknell University Press.
Kincaid, H. (1997). *Individualism and the unity of science: Essays on reduction*. Lanham: Rowman & Littlefield.
Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge: Cambridge University Press.
Latour, B. (1988). *The pasteurization of France*. Cambridge: Harvard University Press.
Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 225–258). Cambridge, MA: MIT Press.
Lave, J. (1991). Situating learning in communities of practice. In L. B. Resnick, J. M. Livine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 63–82). Washington, DC: American Psychological Association.
Law, J. (1991). Introduction: Monsters, machines and sociotechnical relations. In J. Law (Ed.), *A Sociology of monsters: Essays on power, technology, and domination* (pp. 1–23). London: Routledge.
Law, J. (1999). After ANT: Complexity, naming and topology. In J. Law & J. Hassard (Eds.), *Actor network theory and after* (pp. 1–14). Oxford: Blackwell.
Mansfield, N. (2000). *Subjectivity : Theories of the self from Freud to Haraway*. New York: New York University Press.
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183.
Miller, W. B. (1955). Two concepts of authority. *American Anthropologist, 57*, 271–289.
Moore, J. L., & Rocklin, T. R. (1998). The distribution of distributed cognition: Multiple interpretations. *Educational Psychology Review, 10*, 97–113.
Myers, F. R. (1986). *Pintupi country, pintupi self: Sentiment, place, and politics among western desert aborigines*. Washington, DC: Smithsonian Institution Press.
Nietzsche, F. W. (1956). *Birth of tragedy. Genealogy of morals* (F. Golffing, Trans.). Garden City: Doubleday.

Nishigaki, T. (2006). The ethics in Japanese information society: Consideration on Francisco Varela's the embodied mind from the perspective of fundamental informatics. *Ethics and Information Technology, 8*, 237–242.

Osteen, M. (2002). *The question of the gift: Essays across disciplines*. London: Routledge.

Resnick, L. B., Levine, J. M., & Teasley, S. D. (Eds.). (1991). *Perspectives on socially shared cognition*. Washington, DC: American Psychological Association.

Ryle, G. (1949). *The concept of mind*. London: Hutchinson.

Said, E. W. (1985). *Beginnings: Intention and method*. New York: Columbia University Press.

Said, E. W. (2000, December 17). Deconstructing the system (Review of power: Essential works of Foucault, 1954–1984, Vol. 3, by M. Foucault). In *New York Times book review* (pp. 16–17).

Selinger, E., Aguilar, J., & Whyte, K. P. (2011). Action schemes: Questions and suggestions. *Philosophy and Technology, 24*, 83–88.

Smith, H. (1991). *The world's religions*. San Francisco: HarperSanFrancisco.

Star, S. L. (1991). Power, technology and the phenomenology of conventions: On being allergic to onions. In J. Law (Ed.), *A Sociology of monsters: Essays on power, technology, and domination* (pp. 26–56). London: Routledge.

Strathern, A., & Stewart, P. J. (1998). Seeking personhood: Anthropological accounts and local concepts in Mount Hagen, Papua New Guinea. *Oceania, 68*(3), 170–188.

Suchman, L. (1998). Human/machine reconsidered. *Cognitive Studies, 5*(1), 5–13.

Tufts, A. (1996). Eurotransplant to allocate kidneys by computer. *Lancet, 347*(9011), 1326.

Udehn, L. (2001). *Methodological individualism : Background, history and meaning*. New York: Routledge.

Van der Velden, M. (2009). Design for a common world: On ethical agency and cognitive justice. *Ethics and Information Technology, 11*, 37–47.

Verbeek, P.-P. (2006). The morality of things. In E. Selinger (Ed.), *Postphenomenology: A critical companion to Ihde* (pp. 117–128). Albany: SUNY Press.

Verbeek, P.-P. (2009). Cultivating humanity: Toward a non-humanist ethics of technology. In J. K. Berg Olsen, E. Selinger, & S. Riis (Eds.), *New waves in philosophy of technology* (pp. 241–263). Houndmills: Palgrave Macmillan.

Wagner, R. (1991). The fractal person. In M. Godelier & M. Strathern (Eds.), *Big men and great men: Personifications of power in Melanesia* (pp. 159–173). Cambridge: Cambridge University Press.

White, L. (1939). Mind is minding. *Scientific Monthly, 48*, 169–171.

Wood, M. (1998). Agency and organization: Toward a cyborg consciousness. *Human Relations, 51*, 1209–1226.

# Chapter 5
# Some Misunderstandings About the Moral Significance of Technology

**Peter-Paul Verbeek**

**Abstract** The discussion about moral agency and technology is troubled by some severe misunderstandings. Too often, the claim that technologies are involved in moral agency is misread for the claim that technologies are moral agents themselves. Much of the discussion then focuses on the question whether not only humans but also technologies can have intentionality, freedom, responsibility, and, ultimately, moral agency. From the perspective of mediation theory, this discussion remains caught in a dualist paradigm that locates human beings and technological artifacts in two separate realms, humans being intentional and free, technologies being instrumental and mute. Against the question to what extent technologies can be moral agents, mediation theory makes it possible to investigate how intentionality, freedom, and agency are in fact the result of intricate connections and interactions between human beings and technological artifacts. Rather than checking if technologies can meet a pre-given criterion of moral agency, we need to re-conceptualize the phenomenon of moral agency itself in order to understand the roles of technologies in our daily lives.

## 5.1 Introduction

"So you really think we should blame cars for traffic accidents?" I must have heard this question at least once a month over the past years, when I was working on my book Moralizing Technology: Understanding and Designing the Morality of Things (Verbeek 2011). At almost every lecture I gave about the moral significance of technologies, there would be somebody in the audience who could not bear the idea that we should allow material artifacts to play a role in the realm of ethics. "What

P.-P. Verbeek (✉)
Department of Philosophy, University of Twente, Enschede, The Netherlands
e-mail: p.p.c.c.verbeek@utwente.nl

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_5,

about human responsibility?" they usually acclaimed in slight despair. To be followed immediately by arguments like: "If we are going to blame things for evil practices, human beings would have a comfortable place to hide themselves!" And: "If technologies are allowed to influence our moral actions and decisions, we will inevitably develop an undesirable form of moral laziness!" In their view, ethics is an exclusively human affair; if we allow technologies to be part of it, we gamble with the crown jewel of human civilization.

In this contribution I would like to offer some relief to these worries, by addressing some of the most stubborn misunderstandings about the moral significance of things.[1] The core of my argument will be that we need to develop an alternative account of the relations between humans and technologies in ethical theory – an account that allows us to understand how moral practices are coproductions of humans and technologies, rather than exclusively human affairs in which technologies can only play instrumental or obstructive roles. Human beings and technological artifacts have become so closely connected in our everyday lives, that even our moral perceptions and decisions have become technologically mediated. Only by recognizing this interweaving of humans and technologies can we take responsibility for the ways in which technologies have an impact on society and on human existence – in practices of technology design, implementation, and use.

My claim is that the resistance against the idea that technologies are morally significant is in fact a resistance against the need to give up the modernist idea that actions and decisions can only be moral when they are the sole product of individual human choice without external influences. Allowing material objects to play a central role in things we have always considered to be our own domain appears to be too big a hurdle to take. In this sense, there is no doubt that Sigmund Freud would have found the current discussion about morality and technology rather amusing. Even without being a full-blown Freudian, one can probably see the value of his claim that modern science has caused humanity various 'narcissistic offences' (Freud 1955, 137–144). Science humiliates human beings with insights that urge us to replace our all-too-high self-esteem with new forms of humbleness. Copernicus, for instance, showed that not the earth but the sun is the center of the heavens. And Darwin showed that the human being is not the central entity in God's creation, but just a mammal sharing common ancestors with modern apes. To be followed by Freud himself, who demonstrated that it is often not conscious decisions but unconscious factors that shape human behavior.

The fierce resistance against the idea that human morality is interwoven with nonhuman entities shows that these narcissistic wounds have not quite healed yet: the autonomous subject appears to be reluctant to receive yet another blow. When philosophers of technology even hint towards the possibility that technological artifacts might have moral significance, immediately worries arise about human autonomy and responsibility, and even fears that we will end up in a kind of

[1] This contribution incorporates and substantially expands elements of the brief 'reply to critics' I gave in the journal *Philosophy and Technology,* in a symposium on my book *Moralizing Technology* (Verbeek 2012).

pre-modern animism, which brings 'spirit' to things rather than humans. Human dignity itself seems to be at stake when technologies get involved in moral actions and decisions.

These fears are often based on wrong assumptions, though. Let me, therefore, carefully investigate the most important misunderstandings about the moral significance of things, ranging from ideas about their alleged agency to the possibility of technological intentionalities and their implications for human freedom and responsibility.

## 5.2 Do Artifacts Have Morality?

Why would it make sense to speak about technologies in moral terms in the first place? When ethics is so obviously a human affair, why bother giving the nonhumans a place in ethics as well? The reasons for this expansion of ethics are actually quite understandable if one takes into account the profound influence that various technologies have come to have on the decisions and actions of human beings. Navigation systems in cars help us not to exceed the speed limit, antenatal diagnostic technologies inform moral decisions about abortion, telecare devices reorganize relations of care and the moral dimensions connected to it. If ethics is about the question of 'how to act?' or 'how to live?', and technologies help to shape how we act and live, there is good reason to claim that technologies have explicit moral significance.

The question remains, though, how to understand this moral significance. As Pitt in his contribution to this book argues, one can see this significance as purely instrumental. Technologies do not carry any morality in themselves, but are just neutral instruments by which human beings can actualize and implement their morality. In contrast, in their contribution Illies and Meijers attempt to see morality as a human affair that is technologically situated: technologies provide a context in which human beings make moral decisions. The position I will defend here is a bit more radical; I will show that technologies are intrinsically involved in moral decision-making. This does not imply, to be sure, that they are moral agents themselves. But it does imply that moral agency needs to be understood as a fundamentally hybrid affair.

A relevant framework to analyze the moral significance of technology is offered by the postphenomenological approach of 'mediation theory' (Verbeek 2005). This approach studies technologies as mediators between humans and reality. The central idea is that technologies-in-use help to establish relations between human beings and their environment. In these relations, technologies are not merely silent 'intermediaries' but active 'mediators' that help to constitute the entities that have a relationship 'through' the technology.

The paradigmatic example I elaborate in my recent book 'Moralizing Technology' (Verbeek 2011) is antenatal diagnostic technology, such as obstetric ultrasound. This technology is not merely a neutral interface between expecting parents and

their unborn child: it helps to constitute what this child is for its parents and what the parents are in relation to their child. By revealing the unborn in terms of variables that mark its health condition, like the fold in the nape of the neck of the fetus, ultrasound 'translates' the unborn child into a possible patient, congenital diseases in preventable forms of suffering (provided that abortion is an available option) and expecting a child into choosing for a child, also after the conception.

Ultrasound does not force people to have an abortion, obviously. But at the same time we would not do justice to the role of obstetric ultrasound in the moral decisions of expecting parents if we would say that it is morally neutral. By making the fetus present in a highly specific way, obstetric ultrasound substantially informs moral decisions regarding abortion, without determining them. The moral question of 'how to act' gets answered not only on the basis of the input of human beings but also of nonhuman entities. Morality appears to be a coproduction of humans and nonhumans. We do not make moral decisions about abortion as autonomous subjects, but neither are we steered by technology as if we were determined objects. Morality is technologically mediated: it takes shape in technologically mediated relations between humans and reality.

This does not mean that morality is technologically *determined*, though. After all, when we see how moral actions and decisions take shape in interaction with technologies, we can actively intervene in this interaction. Design processes can be reorganized in such a way that they anticipate the mediating roles of technologies. Use practices can be enriched by equipping users with the ability to 'read' how the technologies they use help to shape their actions and decisions, so that they can deal with these mediations in creative and responsible ways.

From the perspective of mediation, therefore, the moral significance of technology is in the technological mediation of morality. By organizing relations between humans and world, technologies play an active, though not a final, role in morality. Technologies are morally charged, so to speak. They embody a material form of morality, and when used, the coupling of this 'material morality' and human moral agency results in a 'composite' moral agency. This implies that technological artifacts should, indeed, be located in the realm of moral agency: moral agency cannot be understood without taking into account how it takes shape through technological mediations.

## 5.3 Do Artifacts Have Agency?

For many critics, the central issue in discussions about morality and technology is the question to what extent technologies can qualify as moral agents or as having moral values by themselves. Joe Pitt's contribution to this volume is a clear example of this position. According to Pitt, technological artifacts can only be instrumental – there is no other way in which they can actively contribute to moral actions and decisions than by facilitating human activities. Morality is in humans, not in things.

Let me first state that it is in fact hard to find scholars who seriously defend the thesis that technologies can be full-blown human agents just like human beings are. To be sure, there are scholars – including myself – who explicitly speak about technologies in terms of moral agency. But it is a true misunderstanding to think that this implies that technologies in themselves 'have' a form of agency that we normally only attribute to human beings. Rather, these scholars propose to reconceptualize the very concept of agency itself, in view of the close intertwinement of human beings and technological artifacts.

The work of Bruno Latour is a good example in this context. For many critics of the 'moral significance of technology' thesis, Latour's 'symmetrical approach' to humans and things is highly problematic, because it raises the suspicion of anthropomorphizing things. Latour refuses to make an a priori distinction between human and nonhuman entities (Latour 1993). Approaching both types of entities with different sets concepts, according to Latour, would make it impossible to conceptualize their interaction adequately. Nonhuman entities, after all, do not only play a role in the material world, but help to shape the social world as well. This becomes especially visible in the moral domain. Objects like speed bumps and door springs embody moral norms: they help us to slow down near schools and to close the door behind us. They are not just neutral instruments that humans can use to realize their own, autonomous intentions: objects help to shape what humans do and even want.

To be sure, human beings do not always need to obey the forces that are exerted upon them. But even when humans ignore or resist the impact of technologies, these technologies can still have a profound moral impact. The introduction of antenatal diagnostic tests, for example, has definitively changed human responsibilities regarding pregnancy. Even the decision *not* to use these tests has become a moral decision now, as some 'wrongful life' cases show in which children sue their parents or doctors for the fact that the were born rather than having been aborted. Even technologies that are not used can have an impact on human morality (Verbeek 2011). Human agency, therefore, and in many cases also our moral agency, has become intertwined with material objects. All those ethicists who are complaining about the moral decay of our society, according to Latour, should learn to include things in their reflections; when these are taken into account as well, the world appears to be choc-a-bloc with morality.

This analysis does not imply, though, that scholars like Latour claim that material objects, 'have' moral agency just like human subjects do. In fact, the very question whether artifacts can be moral agents originates from a very specific metaphysical orientation: the modernist separation of human subjects and nonhuman objects that scholars like Latour intend to overcome. To be sure: Latour does speak about technologies in terms of moral agency. But anyone who is a bit familiar with the metaphysics of Actor-Network Theory will immediately realize that this does not imply that things can be moral agents in themselves. The central idea in Latour's approach, after all, is that no entity can be something 'in itself'. Only in relations to other entities can they become meaningful and relevant: only networks turn entities into actors. Speed bumps can never be 'moral agents' in themselves, but only in relation to human beings whose morality is affected by these things.

Only from a metaphysics in which humans and things are radically separated – humans being active and intentional, nonhumans being mute and inert, as Bruno Latour has elaborated so convincingly (Latour 1993) – it becomes relevant to ask whether things, just like humans, can be moral agents. When we give up this separation, a refreshingly new picture emerges. And this is exactly what happens in various contemporary approaches in philosophy of technology. Such approaches urge us to see moral agency not as inherent in things, but as the outcome of complex interactions between humans and things.

In my own theory of moral mediation (Verbeek 2011) I broaden the perspective by analyzing the moral significance of technologies in terms of their mediating roles in human-world relations. When technologies are used, they do not only help to organize human actions but also our experiences and perceptions. Speed bumps help to organize how we experience the road and how we behave on it; email and cell phones help to organize norms regarding human communication. This phenomenon of moral mediation does not make things moral agents in themselves. Only in the context of the relations human beings have with them can they help to organize people's moral behavior and perceptions. At the same time, though, this implies that technologies do find themselves in the realm of moral agency. Agency is distributed over humans and things, as it were: if one of the two were missing, this type of agency could not exist.

The question remains, though, whether this explanation can be a real assurance for critics of the 'morality of technology' thesis. The argumentation above, after all, does not only imply that things do not 'have' moral agency – the most crucial point is: neither do humans. Morality is a hybrid affair; it cannot be located exclusively in things, but not in humans either. Each in their own way – distinct, but not separated – humans and things contribute to moral actions and decisions. Reducing ethics to an exclusively human affair leaves us with a drastically impoverished world. Because such a 'humanist approach' starts from a radical separation between subjects and objects, it forces us to choose between either reserving moral agency to the human domain or to claim that nonhuman entities can be moral agents as well. In the real world in which we all live, though, such purified subjects and objects do not exist. Actual moral actions and decisions take place in complex and intricate connections between humans and things, which have moral agency as a result rather than as a pre-given ontological characteristic.

Defending the idea that things have moral significance, therefore, should not be understood as a defense of animism but rather as a critique of humanism. Instead of claiming that material objects are 'spirited', scholars who defend the idea that technologies are morally significant move away from ethical approaches that isolate and immunize human existence from its material conditions and contexts. As I will argue below, such a hybrid approach to the relations between humans and things does not reduce human morality, but adds to it; it shows dimensions that normally remain underexposed. Conceptualizing the moral significance of things does not undermine human responsibility by blaming cars for accidents, but rather expands the ways in which we can design, implement, and use technologies in responsible ways.

## 5.4 Can Things Have Intentionality?

When criticizing the 'morality of technology' thesis, typically two aspects of moral agency play a central role. First, moral agency requires intentionality: in order to be held morally responsible for one's actions, you need to have had the intention to act in this specific way rather than having done it accidentally. And second, freedom is required: if someone was forced to act in a specific way, this person cannot be held morally responsible for that action. Both aspects seem crucial for moral agency, and at the same time both aspects seem to be exclusively human.[2] The concept of moral mediation, though, makes it possible to reconceptualize both aspects of moral agency in a 'hybrid' way.

Let me first address the issue of intentionality and technology. In the history of philosophy, intentionality has been conceptualized in two distinct ways. One definition comes from phenomenology, where Brentano and his followers asked attention to the intrinsic directedness of human beings towards their environment. Humans cannot just 'think', they always think *something*. Just like we cannot just hear, see, or feel, but always hear, see, and feel *something*. This directedness is an intrinsic element of the relations between human beings and their world – while these relations can be conceptualized in various ways, e.g. in terms of consciousness (Husserl), or perception (Merleau-Ponty), or being-in-the-world (Heidegger). When technologies mediate these relations, this can have moral implications. After all, technologies then help to shape human interpretations of the world on the basis of which human beings make decisions. Sonograms help to shape our moral decisions regarding abortion, just like warning signals from our navigation system help us decide how fast we drive.

When this occurs, a second definition of intentionality starts to play a role: the human capacity to have intentions, to act purposively. By mediating how we interpret our unborn children or our own driving behavior, technologies help to shape the moral decisions we make. Human intentions, including moral intentions, can be technologically mediated because technologies help to shape our intentional 'directedness' at the world. Sonograms make humans responsible for things they were not responsible for before; it has now become a conscious decision to let a child be born with Down's syndrome, for instance.

Technologies, in other words, give direction to both human experiences and actions. And this is in fact precisely what the Latin word 'intendere' means: to give direction. Intentionality is not an exclusively human affair; technologies find themselves in the realm of intentionality as well.

Some scholars fiercely resist the idea of technological intentionality, though. In order to show how erroneous it is to attribute intentionality to things, for instance,

[2] I am well aware that a proper definition of moral agency requires other elements as well, including the capacity of moral reasoning, as Illies and Meijers argue in their contribution to this volume. But for this capacity the same type of arguments can be developed that I will develop for intentionality and freedom, i.e. that they should be seen as the result of a complex interplay between human and nonhuman entities, rather than a property that both humans and nonhumans can possess.

philosopher of technology Martin Peterson takes the idea that things 'give direction' to humans to the absurd context of the impact that mountains have on humans (Peterson 2012). Of course, mountains have a specific impact on how human beings act and which decisions they make when they climb it, he states, but it would be absurd to see this impact as a result of the intentions of that mountain. Things cannot have intentions themselves, Peterson claims; they can only have an intentional *history*. Only because human beings can intentionally design things in specific ways, things can be the 'carriers' of human intentions. Not the fact that things help to shape practices and experiences makes them morally significant – in that case, non-technological entities like mountains would be moral agents too – but the fact that their impact has its origins in human intentions. Therefore, Peterson concludes that the idea of 'technological intentionality' is "either false or misleading".

Critiques like this show how hard it is to conceptualize the moral significance of things from a radical modernist point of view. The radical separation between humans and technologies make it impossible to see the mediating role of technologies in human decisions and actions. Peterson fights against the idea that technologies can 'have' intentionality, while my claim is simply that, in mediating how humans are directed at reality, technologies help to shape moral intentions.

What is misleading here, in fact, is the reductionist assumption that only human 'input' can make technologies morally significant. The mediating role of technologies in moral actions and decisions cannot be entirely reduced to the intentions of designers and users, after all; some moral mediations emerge without the explicit intention of any human agent. Obstetric ultrasound, again, is a good illustration of this. The technology of ultrasound was not explicitly developed for medical diagnostic purposes, and certainly not to change abortion practices. But as soon as it got to be used to make visible the fetus in the womb, it dramatically changed moral practices and decisions regarding pregnancy. Obstetric ultrasound helped to create new forms of responsibility: while the birth of a child with specific congenital diseases used to be a matter of fate, it now has become a matter of choice. Sonograms translate unborn children in possible patients, congenital diseases in preventable forms of suffering, and 'expecting' into 'choosing'. Expecting parents inevitably have to make a decision about the lives of their unborn children – and also the decision not to use the technology is an explicit choice. The decision whether or not to have an abortion, therefore, is thoroughly mediated by obstetric ultrasound – without anybody having explicitly wanted this situation to occur. To be sure: this does not imply that this morally mediating role of technology is undesirable. Rather, the example illustrates that new technologies always create a new moral landscape in which human beings have to learn to orient themselves.

This example, however, does not only refute the claim that technologies can only have an intentional history; it also shows that limiting morality to human beings makes it impossible to give technologies an intentional future. In practices of design and redesign, of implementation and use, human beings can actively engage with the moral significance of technologies. Rather than taking away morality from human beings, addressing the moral significance of technology actually adds to it, enabling designers and users to anticipate, assess, and design moral mediations in technology.

Current technological developments, to add to this, bring about an even more complicated relation between technology and intentionality than present in the ultrasound example. Some contemporary technologies cannot simply be 'used' anymore, but start to merge with our physical environment and with our own bodies. Ambient Intelligence technologies, for instance, result in 'smart environments', that actively register and monitor events, and react and intervene accordingly. Smart beds in geriatric hospitals are good examples here; they detect if patients fall out of bed or step out of their beds. When such smart environments are explicitly designed to influence people's decisions and behavior, as is the case with so-called 'persuasive technologies', they embody a truly new form of moral significance. An interesting example here is the 'persuasive mirror'. This mirror, which is in fact a flat screen monitor with a built-in camera, is designed to persuade its users to adopt a healthier lifestyle by presenting them with an image of how they will look in the future if they would stick to their current pattern of living (Knight 2005). The intentions of people using technologies like this are not so much 'mediated' as they are 'induced'. They are not the result of 'using' a technology – rather, technologies use human beings here to do their work.

At the other extreme, technologies do not merge with our environment but with our bodies. Prostheses are good examples of this, especially when they are connected to our nerve system, as is the case ever more often. When it comes to the moral significance of these 'mergers' of humanity and technology, brain implants are good examples as well. The technology of Deep Brain Stimulation (DBS) is rapidly gaining popularity in the treatment of neurological and psychiatric conditions like Parkinson's disease, deep depressions, or obsessive compulsory disorder. By bringing in an electrode deeply into the brain, specific parts of the brain can be activated. This can, for instance, dramatically reduce the motor symptoms of Parkinson's disease.

Such implants, however, often also have an impact on people's character. A well-known side effect of DBS for Parkinson, for instance, is that patients can start to develop uninhibited behavior. The Dutch medical journal Tijdschrift voor Geneeskunde discussed the case of a patient suffering from Parkinson's disease whose condition improved markedly after having been implanted (Leentjens et al. 2004). But while the symptoms of the disease were reduced, his behavior developed in an uninhibited way. He got involved in extramarital relationships, spent too much money, and did not have real awareness of his behavior change until the DBS was switched off for medical reasons. But at that moment his Parkinson's symptoms returned; again, he was entirely bedridden and dependent. There appeared to be no middle way; he would have to choose between a life with Parkinson's disease, bedridden – or a life without the symptoms, but so uninhibited that he would get himself into continual trouble. Eventually he chose – with the DBS switched off – to be admitted to a psychiatric hospital, where he could switch the DBS on and suffer fewer symptoms of the disease, but where he would also be protected against himself.

In this case, moral intentions – for instance regarding adultery, dealing with money, et cetera – are not so much mediated by technologies. Rather, the intentions

are the product of a hybrid entity, half human, half technology. Blurring the physical boundaries between humans and technologies also results in the blurring of intentional boundaries.

Intentionality, then, has a complex relation with technology. The conclusion that we need to reserve the concept for humans simply fails to do justice to the manifold ways in which human intentions are intricately connected to technologies – ranging from mediated and induced intentions to the fully-fledged hybrid intentions of human-technology assemblies like people with brain implants.

## 5.5 Can Freedom Be Technologically Mediated?

But what about freedom – the second requirement for moral agency that I mentioned above? If freedom is required to qualify as a moral agent, how could we possibly include things in the realm of moral agency?

Again, the central idea in the approach of moral mediation is not to attribute freedom to things, but to include the mediating role of things in our notion of freedom. From a radical modernist point of view, it is quite a challenge to consider technologically mediated actions as forms of moral agency. After all, if our behavior is influenced by technologies, we cannot consider this behavior to be the result of autonomous decisions anymore. Can we call it a moral action if somebody slows down near a school because there is a speed bump on the road? This question is only relevant, though, if freedom is understood as negative freedom, to use Isaiah Berlin's concept: the absence of external influences (Berlin 1979). Such a conception of freedom is only meaningful if human beings are understood as autonomous subjects, living in a world of external objects. When human beings are understood in terms of their relations to the world, though, this concept of autonomy becomes too narrow.

Isaiah Berlin's concept of 'positive freedom' is much more relevant here: freedom is not freedom-from but freedom-to. It is not the absence of constraints, but the presence of the capability to act. From this viewpoint, mediating technologies do not take away moral agency, but rather are its basis. Obstetric ultrasound does not force parents to have an abortion when they are expecting a child with Down's syndrome; rather, their moral agency comes about in the way they develop a relation to obstetric ultrasound. Human beings are no helpless victims of mediating technologies. We can get actively involved in how these technologies have an impact on us. By critically examining how technologies help to shape situations of choice and frameworks of interpretation, it becomes possible to take responsibility for one's technologically mediated agency.

In fact, it is the denial of the mediating role of technologies in human freedom that makes people not free. Only when making the mediations explicit, we can develop a free relation to it – understood as positive, not as negative freedom. A free relation to technology does not require the absence of its influence, but the presence of the ability to be actively involved in the way in which one is constituted as a

moral subject. In these forms of active involvement, we can take responsibility for our technologically mediated existence.

Yet, mediated responsibility is a problematic concept as well, from a radical modernist approach. In their contribution to this volume, for instance, Anthonie Meijers and Christian Illies claim that my position would have the absurd consequence that we should put part of the blame for a murder and a computer hack on the pistol and the computer:

> In the case of a man using a pistol Verbeek would argue that the two form an association and that the man-pistol association has moral agency and is accountable. The association as such becomes blameworthy. That however, blatantly contradicts our practice of blaming and punishing. We do not (and we should not!) put the murderer *plus* his pistol, or the hacker *plus* his computer, in prison. In such cases it is the human agent alone who, according to standard moral practice, is blameworthy. (Illies and Meijers, this volume)

From the perspective of mediation theory, however, this is quite a remarkable statement. Recognizing that pistols and computers help to shape our moral actions and decisions, after all, does not imply that it should be possible to *blame* them for their mediating roles. Rather than to keep asking the question whether artifacts can be agents just like humans can, mediation theory simply recognizes that humans and artifacts can have distinct roles in the constitution of moral agency. From the perspective of moral mediation it is perfectly possible to acknowledge that one's moral actions and decisions are technologically mediated without giving up the possibility to take responsibility for these actions and decisions. Technological mediations, after all, do not make human beings powerless. On the contrary: they make it possible to live our lives in specific ways, while we also have the ability to develop an active and critical relation to these mediations. Nobody has to choose to have an abortion when an ultrasound scan reveals a serious disease. Still, the mere possibility to have a scan inalterably conditions our existence: we now have to make a decision.

The claim that moral agency is a hybrid affair does not imply that things are moral agents just like humans are. While agency cannot be limited to humans, acknowledging that things have a share in moral agency does not make them moral agents. In fact, the very possibility to take responsibility is one of the main reasons to take the role of things in moral agency very seriously. One the one hand, we do not take responsibility in a vacuum but in a thoroughly mediated situation, as the ultrasound example shows. Sonograms do not 'act' on themselves, but nevertheless they fundamentally shape what we can feel responsible for and how we can take on that responsibility. And what is more, acknowledging this moral role of technologies makes it possible to take responsibility for it, and to help shape it in practices of design, implementation, and use. We can only reorganize practices around technologies when we understand the precise role technologies have in them. Seeing the moral significance of technologies makes us more responsible, rather than less.

This does not imply, to be sure, that human beings can always take full responsibility for the mediations that eventually result. After all, the contextual and relational approach to technologies implies that we should not overestimate our possibilities to organize and design the moral significance of things. We always need to

recognize the fact that all technologies are multistable, in Don Ihde's words (1990). Technologies often end up in different relations with human beings than their designers expected, and therefore their mediating power is hardly predictable.

But this unpredictable character of technological mediations does not make human beings entirely powerless. Industrial Designers, for instance, have developed various methods to make an educated guess about possible use practices. And once we see the phenomenon of mediation, including its moral dimensions, it is our moral responsibility to make such an educated guess and to design 'for the good'.

## 5.6 Conclusion: Is There a Symmetry Between Humans and Technologies?

The central theme in discussions regarding the moral significance of technologies has proven to be the question to what extent we can really attribute aspects of moral agency to technologies. While some scholars fear that we throw out the child with the bathwater when we deny technologies this role, others feel that it goes way too far to claim that technologies have moral capacities similar to humans.

Behind all of these discussions, there seems to be a common misunderstanding. This is the alleged 'symmetry' between humans and nonhumans. This concept of symmetry, which gives up any a priori distinction between humans and nonhumans, originates in Bruno Latour's work. By claiming that we should analyze humans and nonhumans in symmetrical ways, Latour aims to make it possible to see the continuity between humans and nonhumans, rather than taking the distinctions between them as a starting point. And seeing this continuity is needed to be able to understand how nonhuman entities do not only play a role in the material world, but also in the social world. Interpreted radically, such a symmetrical approach implies that not only humans but also things 'act' – and from there it is only a small step towards defending that things can be moral agents as well.

Yet, this symmetry is not essential for conceptualizing the moral significance of technology. And this is in fact what distinguishes my own 'post-phenomenological' account of technology from Actor-Network Theory, despite the many forms of kinship that are there as well. Symmetry is what one gets when using a mirror: a mirror image is completely symmetrical to its original, and it derives all of its main characteristics from that original. In my approach, however, there is no symmetry, but interaction and mutual constitution. Things are not symmetrical to humans, but together, humans and things constitute myriad 'hybrid entities'. In this approach, it remains very relevant to make a distinction between humans and things – it is not the distinction between humans and technologies that we need to depart from, but their radical separation (see also Verbeek 2005, 166–168).

This subtle difference between 'separating' humans and nonhumans on the one hand versus making a 'distinction' between them on the other, makes all the difference in discussions about the moral significance of technologies. The central question in these discussions should not be: can technologies be moral agents just like

humans are? Rather, what really matters is: what role do technologies play in morality? And once this is the question we aim to answer, we can see that there are various ways in which technologies help human beings to answer moral questions and to behave in moral ways. And, what is more: understanding this morally mediating role of technologies makes it possible to deal with it in a responsible way, in practices of design and policy-making.

When we stick to a modernist separation or isolation of subjects and objects, technological mediations of morality are at odds with the autonomy of the moral subject, and can never be seen as full-blown elements of moral agency. But giving up this metaphysical 'apartheid' should not bring us to the other extreme of denying all distinctions between both poles. Morality is neither to be found in the objects themselves, nor in autonomous subjects. It only comes in relations between subjects and objects, where objects have moral significance and subjects are engaged in mediated relations with the world.

This subtle rearrangement of the relations between humans and nonhumans brings us back to discussions about the Enlightenment, when the subject-object distinction started to be a central theme in philosophy. In his famous lecture What is Enlightenment, Michel Foucault discusses how, for Immanuel Kant, Enlightenment meant "a way out of immaturity". He defined immaturity as "a state of our will that makes us accept someone else's authority to lead us in areas where the use of reason is called for." Some of the objections against the moral significance of technology seem to imply a similar logic. It seems that we have to make a choice between using reason or letting things decide for us what to do. And this mirrors what Foucault, in the same text, calls 'the blackmail of the Enlightenment': if you are not entirely with it, you are against it (Foucault 1997).

From the perspective of moral mediation, the opposite is true. Maturity in our thinking about technology requires that we no longer exclude technologies from the realm of ethics. Only by acknowledging the fundamentally mediating role of technologies in moral actions and decisions can we better understand the character of human morality. And, more importantly, can we take responsibility for the material world in which we live our lives.

## References

Berlin, I. (1979). Two concepts of liberty. In *Four essays on liberty* (pp. 118–172). Oxford: Oxford University Press.

Foucault, M. (1997). What is enlightenment? In P. Rabinow (Ed.), *M. Foucault, ethics: Subjectivity and truth*. New York: New Press.

Freud, S. (1955). A difficulty in the path of psycho-analysis. In J. Strachey (Ed. & Trans.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 17). London: Hogarth Press.

Ihde, D. (1990). *Technology and the lifeworld*. Bloomington: Indiana University Press.

Knight, W. (2005, February 5). Mirror that reflects your future self. *New Scientist*, (2485), 23.

Latour, B. (1993). *We have never been modern* (C. Porter, Trans.). Cambridge: Harvard University Press. (Translation of: *Nous n'avons jamais été modernes*, Paris: La Découverte, 1991).

Leentjens, et al. (2004). Manipuleerbare wilsbekwaamheid: Een ethisch probleem bij elektrostimulatie van de nucleus subthalamicus voor ernstige ziekte van Parkinson. *Nederlands Tijdschrift voor Geneeskunde, 148*, 1394–1398.

Peterson, M. (2012). Three objections to Verbeek. In E. Selinger et al. (Ed.), *Book symposium on Peter Paul Verbeek's moralizing technology: Understanding and designing the morality of things, philosophy and technology* (Vol. 25, pp. 619–625).

Verbeek, P. P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. University Park: Penn State University Press.

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. Chicago/London: University of Chicago Press.

Verbeek, P. P. (2012). The irony of humanism: On the complexities of discussing the moral significance of things. In E. Selinger et al. (Ed.), *Book symposium on Peter Paul Verbeek's moralizing technology: Understanding and designing the morality of things, philosophy and technology* (Vol. 25, pp. 626–631).

# Chapter 6
# "Guns Don't Kill, People Kill"; Values in and/or Around Technologies

**Joseph C. Pitt**

**Abstract** This paper presents a defense of the Value Neutrality Thesis with respect to technological artifacts. While it may be case that people build artifacts with certain ends in mind – the values of the people doing the building are not in the artifacts. Why this is so is a function of three things: (1) lack of empirically identifying characteristics of values and (2) an endorsement of a pragmatic conception of values as motivators of human action, and (3) a conception of decision-making that necessarily includes values.

## 6.1 Introduction

I first encountered the slogan "Guns don't kill, people kill" when driving around the drill field at Virginia Tech in the late 1970s and it has bedeviled me ever since. The bumper sticker was issued by the National Rifle Association as part of a campaign to secure the right of individuals to own firearms. I am not going to worry the issue of whether or not the second amendment to the U. S. Constitution gives individuals the right to own weapons. What I am going to do is examine the question of whether or not technological artifacts have, in some sense of "have", values. This is because there is both something right about what the slogan says and something wrong. What's right is that by itself guns don't do anything. On the other hand, humans rarely kill anything with their bare hands and so, it is humans with guns that do the killing, not humans by themselves. Nevertheless, I am led to the conclusion that in fact it is people who do the killing, not the gun. That guns are used to kill is not in dispute.

J.C. Pitt (✉)
Department of Philosophy, Virginia Tech, Blacksburg, VA, USA
e-mail: jcpitt@vt.edu

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_6,

The basic position I will defend is the value neutrality thesis.[1] The value neutrality thesis (VNT) states:

> Technological artifacts do not have, have embedded in them, or contain values.

There are authors such as Langdon Winner who want to argue that certain technologies embody the values of an individual (e.g., Robert Moses in the Long Island Expressway case) or a power elite (nuclear energy plants in particular and the electric grid in general). I, instead, want to maintain that the technologies themselves cannot in any legitimate sense embody values. Rather, it is people who have values. This is not to deny that specific technologies may result from individuals attempting to implement their value systems in certain ways. It is a result of recognizing that values are the sorts of things that inanimate objects cannot possess, embody, or have. It is that point I elaborate below.

The structure of the paper is roughly this: first I try to define values in a non-question-begging way, in the process identifying a potential dilemma. I then look at the sense in which all human decision-making is tainted by values. I conclude by distinguishing between a hard and soft version of VNT – in so doing avoiding the dilemma and de-reifying values.

## 6.2 A Potential Dilemma

From the start there arises an apparently insurmountable problem: whether we want to defend or deny the value neutrality thesis, we have to have an account of values. The problem is that I do not believe it is possible to develop such an account without begging the question, i.e., that values are the sorts of things that only humans have. And if this is true, then we have the beginnings of a potential dilemma: on the one hand, if values are the sorts of things that only humans have, then technologies cannot have values by definition. If, on the other hand, values are the sorts of things that humans have, then everything humans do and make, i.e., all our technologies, are, in some, as yet to be defined sense, at least tainted by human values. The first horn of the dilemma leaves us with the issue of the factors involved in making decisions about which technologies to make and how to employ them. The second renders the value question trivial.

## 6.3 Defining "Values"

To turn to the argument I will now develop in some detail, let us begin by asking the question "What are values?" The trick is to not beg the question. If all I did was to assert, as I did above, that values are not the sort of thing artifacts can have, then I

[1] As is my custom, I eschew talking about Technology with a capital "T", favoring directing our attention to specific technologies. See Pitt 2000.

win by default. That is also cheating. Rather than simply stipulate, the case needs to be made for why values are the sorts of things artifacts cannot have in any *meaningful* way. The final conclusion is going to be the claim that so many values are involved in the creation of an artifact that we might as well say it is value neutral.

The primary problem here is the lack of efforts by philosophers to define "value". There is a lot of value talk. Distinctions are drawn between intrinsic and extrinsic values. There is talk of value theory and value judgments. There is also a lot of talk about valuation. As noted already, there is also talk of values being embedded in things. We talk about the values of a society, or a group or a person and sometimes we even list them, but there is little talk of what makes a value a value. The Stanford Encyclopedia of Philosophy doesn't even have an entry for "value" but it does for value theory. But, never to flinch, there is a definition offered on the web page of the Boy Scouts of America:

> Values are those things that really matter to each of us … the ideas and beliefs we hold as special. Caring for others, for example, is a value; so is the freedom to express our opinions. (http://pinetreeweb.com/values.htm)

As pithy as that may sound, it is at least an effort and it has some merit. And in some respects it echoes C.I. Lewis' (1946) thinking. As seems to be the case with most of the philosophers I have consulted, he doesn't address "value" directly, but instead talks of "the objective value-quality in existent things". This he takes to be revealed in the truth or falsity of "predictions of a goodness or badness which will be disclosed in experience under certain circumstances and on particular occasions." (365) As a pragmatist, he understands value in terms of how actions turn out. The predictions of the results of certain actions being true or false then make it possible for him to turn value talk into empirical talk. It is a nice sleight of hand, but a slight of hand, nonetheless. And it doesn't give us a satisfying sense of the meaning of "value".

To attempt to move the discussion further, I propose a different pragmatist account: a value is an endorsement of a preferred state of affairs by an individual or group of individuals that motivates our actions. To some this may sound like a definition of "goal". However, values differ from goals in that goals are objectives to be obtained. Once obtained, a goal is no longer a goal; it becomes something achieved. Values, on the other hand, as a motivation to achieve a preferred states of affairs, serve as action initiators, directing what we do in one direction rather than another. The preferred state of affairs serves as a regulative ideal. Thus, to be an honest person is a preferred state of affairs for most people and remains so even if sometimes we are less than honest. That we are sometimes less than honest does not mean we have given up honesty as a regulative ideal. Circumstances intervene. Furthermore, despite their differences, values, as action motivators, and goals can be rank ordered and those orders can change.

Consider the dying grandmother case: Your grandmother is very close to death. You are her favorite grandchild. She has always wanted you to marry Sally, a lovely girl who lives two houses away. You like Sally, but are in love with Mildrid and unbeknownst to your grandmother have proposed marriage to Mildrid. As her final

request, your grandmother asks you to promise to marry Sally and make a dying old lady happy. You decide to tell your grandmother you will marry Sally, having no intention to do so, but you have insured that grandma dies happy.

Only but the most severe of moral philistines, such as a devout Kantian, would insist you tell grandma the truth. Allowing your grandmother to die happy, as a preferred state of affairs, seems so much more important than telling her the truth at this point. Your decision here is based on thinking beyond yourself. On the other hand, had grandma been perfectly healthy with no signs of early demise on the horizon, you would have told her the truth: that you are marrying Mildrid. Acknowledging that values can shift in their relative rankings does not mean that you are an unrepentant relativist. Or even if you are, you would be a harmless relativist. What is important here is that you have a set of values that are consistent and realistic. That they shift in their relative rankings is just a fact of life. That we are motivated to do one thing rather than another depends on which preferred state of affairs is more important at the time – making your grandmother happy as she dies or rigidly adhering to your desire to be honest. More important still is that in making your decision you did not make it on the basis of whether you would be morally degraded in some way. Rather, your decision was based on what was best for your grandmother, specifically your grandmother. And that may be the heart of the matter. Morality is a social virtue – your values mainly concern how you interact with others for their well being.

## 6.4 The Spectre of Moral Relativism

The standard worry about the view endorsed above is that if values can change, aren't we doomed to relativism? Morals, after all, are supposed to guide our actions. If we are committed to relativism then doesn't that leave us without a general standard of good and evil and hence able to do whatever we want, values and morals be damned, thereby leaving us without a guide as to how we should act in any given situation? If you are a relativist that may be true, which is another mark against relativism. More fundamental, however, is the question of how morals are supposed to guide our actions. Just laying down a moral code with a list of does and don'ts isn't going to do the trick. The world is much too complicated for us to assume that that list is adequate to achieving whatever goal a moral theorist sets. And that perhaps is the problem. The goal of moral theory can't just be internal – to turn you into a good person. This is primarily because being a good person entails how you interact with those around you. It is what you *do* that is important. Without concerns that look beyond the individual actor acting in apparently isolation there is no assistance forthcoming when we face decisions like telling Grandma the truth on her deathbed.

So I endorse Aristotle's view that the goal of moral theory is help us achieve The Good Life – and achieving the Good Life cannot be solely about my life – but the Good Life for all – not just for me, but for all of us. This is not a new thought, just a reminder of what we should be concerned about. Such a goal provides us

with the kind of leverage we need in tricky cases like Grandma's. In the light of what we are actually faced with in terms of challenges to our moral sensibilities, the much-to-be-feared-relativism that so many moral philosophers hold up as anathema seems silly and an ethics that ignores what we do is irrelevant. The job is **not** to develop a high-minded set of principles that cannot be used in the real world situations in which we find ourselves. By accepting the fact that values can be reprioritized in the light of the changing bigger picture puts us in a position to argue that since morals are a particular form of value instantiation, that morals can also be shifted around to accommodate the difficulties we encounter when attempting to work our way toward The Good Life. The point about the Good Life is that when we contemplate achieving it, we have to factor in the fact that the other six billion people on the planet are trying to achieve the same goal. Two things here: first, while we are all seeking the Good Life, it does not follow that we all have the same conception of the Good Life. Second, it is surely true that "no man is an island unto him or herself". We live with others and must take them into account. Further looking beyond my wellbeing entails also looking beyond the consequences of a single action. That is, we need to look at the consequences of the consequences of our actions. Having a consistent, if movable, set of values is one piece of the puzzle of how to achieve that.

## 6.5 Values and States of Affairs

So let us assume for the moment that the proposed account of values as motivators to achieve preferred states of affairs is a viable starting point.[2] But that may be a bit hasty since it already makes values the sorts of things people have and artifacts don't since it make little sense to speak of artifacts having motivations. But that may not be too bad a result. Delaying a defense of this view until later, I want to argue that all human decisions are value-laden and that since any artifact will be the result of many decisions, many values will be involved, so many in fact that it becomes impossible to identify the one value that an artifact embodies, were artifacts to embody values. At the moment, our next step is to clarify the sense in which values are motivators to achieve preferred states of affairs.

A preferred state of affairs is a goal to be achieved. It is not enough to say or write that, for example, eliminating poverty is our goal. Articulating such an objective is important, but unless you do something to achieve that goal, it is merely empty rhetoric. Endorsing the goal means *acting* in such a way as to bring it about, this is the pragmatism part.

[2] Truth in advertising requires that I confess to my philosophical proclivities. I am both a Peircean Pragmatist methodologically and a Humean morally – we cannot address abstract philosophical concepts without a commitment to some view of human nature and the consequences of that view for our understanding of why we do what we do (Pitt 2005). In this I believe Hume is correct when he says that "Reason is and ought only to be the slave of the passions (Hume 1738/1978, p. 425)."

## 6.6 Whose Value?

Values have something to do with states of affairs that are important to humans. It is, however, difficult to determine whose value, whose endorsement, a particular artifact embodies, if, in fact, values are embodied in objects. Here at Virginia Tech some argue that the football stadium instantiates a value. But, and this is an important question: whose value? Here are some possibilities: (1) Football is important to the President, the alumni and the students, especially since if we have a winning team; it accords the university some sort of prestige. (2) The stadium is emblematic of the University – thus those who are proud of the university and being part of it in some way, see the stadium as a symbol for all that is good about Virginia Tech. (3) The stadium stands for the aspirations of the football players who have dreams of playing professional ball and making lots of money. All three of these values, however, are values of people. They may see the stadium as symbolizing their own values, but that doesn't mean the values are in the stadium. But if the values are motivators to achieve certain goals, as we have been arguing, then we can see how they play out. And in seeing how they play out we can determine if they are values on which should we act. Thus, attending or working at a prestigious university should be beneficial in the long run for both students and faculty. And being president of one certainly accords the president a certain status. But if the university acquires its prestige by acting so as to develop a good football team at the expense of high academic standards or supporting faculty research, then it is not clear that the stadium embodies a good value. The same is true for (2) and (3) above. Since universities are educational institutions dedicated to the education of the young and increase in general knowledge, it seems that a football stadium is a poor choice to symbolize the university. And we will have failed in our role as teachers, if the athletes who attend Virginia Tech do so only to make it big in professional sports.

## 6.7 Artifacts Embodying Values

Let us now look at what it would mean for an artifact to embody human values. The problem, of course, lies in the lack of discerning empirical properties of values. We don't know what they look like (hence the pragmatist move to actions – if you claim you have a certain value, then you must do something to show that you, in fact, endorse that value). In Langdon Winner's examination of the construction of the Long Island Expressway (LIE), he claims that it was Commissioner of Parks and Recreation Robert Moses' desire, i.e., his values, to keep inner city residents from the beaches of Long Island that resulted in the overpasses over the LIE being of such a height that city buses, the only mode of transportation available to inner city residents who wanted to visit the beaches, could not pass under them.[3] So Moses' values are embedded in ….what? Are they to be found in the design, i.e., the working drawings, of the

[3] Never mind that Winner's history is in dispute. See Woolgar and Cooper 1999.

LIE? Where would we see them? Let us say we have a schematic of an overpass in front of us. Please point to the place where we see the value. If you point to the double-headed arrow with the height of the overpass written in, you have pointed to a number signifying a distance from the highway to the bottom of the underpass. If you tell me that is Robert Moses' value, I will be most confused. There are lots of numbers in those blue prints. Are they all Moses' values or intentions? Some have to do with other features of roads, such as the depth of the roadbed. How do we differentiate the height of the overpass from the depth of the roadbed in a principled fashion as a human value and not arbitrarily? Similarly, if an engineer claims that this design for an automobile incorporates his value of efficiency, where do we locate that? If he is speaking of aerodynamic efficiency, we can look at the lines of the design, but, that is the drawing representing what the automobile might look like, not the automobile itself. If we look at the automobile itself, can you distinguish the design of the machine from the parts? The fender has a certain curvature that is made possible by the materials out of which it is made – were the materials in the engineer's account of the efficiency of his design? I think not. Likewise for the LIE – if we look at the actual physical thing – the roads and bridges, etc. where are the values? I see bricks and stones and pavement, etc. But where are the values – do they have colors? How much do they weigh? How tall are they or how skinny? What are they?

## 6.8 Technological Values as Sui Generis

Now what is clear is that *if* Robert Moses had certain prejudices and ordered a design for the LIE that facilitated his desire to keep inner city folk from the beaches, then we might conclude it is Moses' intentions, desires, i.e., values, that put certain structures in place, but that does not mean that the structures themselves have his values. If it turns out that because of the low overpasses some bus designers also saw an opportunity to build and sell lower, more economically designed buses, it does not follow that Moses' values are in the better bus design. The line from intention or value to artifact is not as direct as Winner would have us believe.[4]

Essentially, I am arguing that we can't locate the values that are supposed to be embedded in the artifact. If we can't locate them, then it is merely *metaphorical* to say of an object that it embodies human values. This argument is similar to the one I raised against Davis Baird's claim that scientific instruments of a certain kind, spectrometers, embody human knowledge (Baird 2004). When I examine the machine, I can see how it works, what it does, etc., but I can't see the knowledge. Baird's claim is, I think, that the machine itself is the knowledge. There is a certain sense to that, if we buy the other part of Baird's claim, which is that our various accounts of propositional knowledge do not exhaust the forms of human knowledge. If the cash value of that claim is merely to make room for artifacts as embodiments of human knowledge, then it seems he has begged the question.

[4] I have been using "values" and "intentions" somewhat interchangeably here, but only because they seem to indicate the same sort of thing – motivators.

Likewise, the same logic doesn't work for our values problem, for to say that there are more kinds of values than the values that humans espouse, is to beg the question. It is to say the values embedded in technological artifacts are *sui generis* that is, they are of their own kind, which is to say there is something embedded in a object which is a value, but now it is not a human value, it is a technological value – but that needs work and it is not the same as saying that human values are in technological objects. It is to say that technological objects contain technological values. Now if we only knew what technological values looked like we might be getting somewhere – but we don't, so we haven't.

## 6.9 Values and Consistency

Above I argued that values are not goals and that sets of values ought to be consistent and rank orderable, where the rank order can be adjusted. At first blush it is not clear what it means for a set of values to be consistent. If bravery and honesty are consistent does that mean bravery and dishonesty are inconsistent? To decide that we need to know the answer of "consistent with respect to what?" Let me suggest that we judge a set of values to be consistent when they all contribute to the achieving of the Good Life and do not negate the goal or undermine those efforts. Thus, if we agree that a world in which hunger has been eliminated is part of our conception of the Good Life, then the development and use of pollutants that decrease the fertility of the soil is counterproductive – hence inconsistent with the value of eliminating hunger.[5]

Looking again at the problematic that initiated this discussion, "are technologies value neutral?", we have found that locating whose value is embedded in the artifact is very difficult and locating where the value is in the artifact is equally difficult. And these difficulties stem from our lack of identifying characteristics of values such that we could locate them *in* things. We could purse that line of attack: find the empirical characteristics of values such that we could see them or measure them. But then again, maybe we have been asking the wrong question, Let us attempt, at this point, to see what the question is really asking.

## 6.10 Turn the Question Around

Why would anyone deny that technologies are value neutral? One answer that presents itself is "To escape responsibility for their actions." To say that means something like this: the machine made me do it. And that is totally absurd. Machines don't make you do anything. That is the truth in the bumper-sticker "Guns don't kill, people kill". You choose to use this machine to commit that act. You can't blame the machine. Let

[5] Part of this discussion, were it to be extended, should be an examination of long term versus short term consequences of our actions and how they impact our quest for the Good Life. But that would take us far afield.

us take a different example. Perhaps you are driving down the street and a small child suddenly darts in front your automobile. You slam on the brakes, but they fail; and you hit the child – blame it on the machine – or part of the machine, the brakes. Wrong again –for several possible reasons. Either (1) you failed to get the brakes inspected on a proper schedule and keep them properly maintained, or (2) the brake materials or construction was faulty. In the first case the accident is your fault. You were negligent in your upkeep of the automobile. In the second case, whoever installed and/or inspected the brakes failed to do their job. In both cases it is the person who is responsible for the problem, not the machine. We can multiply the examples, but the fact of the matter is that you chose to use that machine and you and the individuals who designed and made that machine/artifact are responsible for its proper function.

A second possible reason to deny that technologies are not value neutral is ideological. That is to say that you are pushing an ideology that advances the claim that a certain power structure or organization is responsible for certain alleged results. Thus Winner, in his *The Whale and the Reactor*, argues that large electric companies in the form of their boards of directors conspire to disenfranchise ordinary citizens. So here, for Winner, we have electric companies, with the compliance of the government, building large nuclear facilities to generate electricity, violating the land, and excluding the people from the decision-making process. In this case the technology is the electrical-industrial complex and one of the values it allegedly embodies is the capitalist drive for making money at the expense of the little people. But this is little more than conspiracy theory run amok.

Anyone who wants to can see the Dark Side of the Force behind any perceived evil. That does not make what-is-perceived-to-be-evil evil nor does it make it the result of the Dark Side of The Force. The trouble with ideological critiques is that their claims are unfalsifiable. If you can’t find the villain, it is because the villain has hidden all the evidence, don’t cha see?! Another tactic of ideologists when faced with failure is to blame the failure on a variety of other factors. Thus, it is not that Marxism is flawed, it is that the particular instantiation of it we saw in the Soviet Union was corrupted by individual ambition.

We started this discussion by laying out a potential dilemma – either values are the sorts of things that humans have or technologies are so value-laden that claiming technologies are embedded with human values is trivial. We have at least examined the first part of the dilemma and discovered that it is, at best, very difficult to find values in things. And, when we speak of humans possessing values, the best that I can do is look to what they do to achieve a professed desired state of affairs. In either case, our tentative conclusion is that humans have values and things don’t. So let us now turn to the second horn.

## 6.11 Many Values

The truth is that while there are many values involved in the creation of an artifact, it does not follow that the artifact is value-laden in any interesting sense. When we understand the process taking us from idea to design to manufacture to marketing

and sales, we find multitudes of value-laden decisions at every step. There are too many to single out any one without a non-arbitrary selection process and we have seen how difficult that is to do. Look at the F-16 fighter jet. It is truly a marvel. It is not just fast, but maneuvers like a dream. It is a technological delight, full of all kinds of neat technological gizmos. The United States military authorizes fly-overs at virtually all home football games at my university because we have a large military training program and many of the pilots are Virginia Tech graduates. It brings delight to all in this case. It is a "WOW!!!" inciting machine. And it can be used to kill. How it is to be used, to delight or to kill, is the result of decisions humans make. Further, what it is *as a machine* is the result of decisions humans made.

The crux of the matter is that all human decisions are value-laden. This also means that all technologies are created by humans making value-laden decisions. What do I mean by the claim that all human decisions are value laden? I will back into the argument supporting this claim by looking first at a long neglected paper by Richard Rudner, "The scientist *qua* scientist makes value judgments" (Rudner 1953). Rudner begins by drawing a distinction between two types of values, epistemic and non-epistemic. I call the latter aesthetic values and place moral, political, social, etc. under that broad category.[6] For epistemic values Rudner has in mind truth, provability, explanation, evidence, testability, elegance, etc. The argument itself is quite elegant. Scientists make choices at every stage of their investigations. For example, to choose between which of two hypotheses to test, the reasoning usually employs such notions as "which of these will we be able to empirically test. If one hypothesis is deemed untestable, the other will be selected unless it too is untestable." That a hypothesis be testable is an important value in scientific inquiry. It is not something written in stone, but rather serves as one among many motivators that guide inquiry. The set of epistemic motivators (together characterizable as the desire to know) that characterize scientific decision-making constitute the values that are constitutive of scientific inquiry as opposed to other forms of inquiry.

The important thing is that at every decision point values are employed. Rudner's contribution here is the recognition that truth and testability and the others are epistemic *values.* "Truth" had never been considered a value prior to this point – just what it was supposed to be is unclear. But by indicating a category of epistemic values, Rudner made it possible to understand science as a more fully human activity than it has been characterized by, for example, the logical positivists.[7] By "more fully human" I mean to imply that it helps to explain why science is not infallible. Scientists make value judgments. Values are motivators to achieve endorsed states of affairs, but they are not objective in any hard and fast way. Thus, when the positivists finally realized that the requirement that empirical claims be verified was too strong a demand, since complete verification is never possible, verification was weakened to confirmation. This was, from this perspective, a change in values and

[6] The position I would like to elaborate but which will take us too far afield is that moral behavior is a form of aesthetics.

[7] It may be seen as something of an irony that Rudner's teacher was Carl Hempel, a student of Rudolf Carnap, one of the founders of positivism.

it had ramifications for the relations among the rest of the epistemic values associated with being scientific. If confirmation of a hypothesis was now the objective, then a confirmed hypothesis, i.e., one deemed acceptable, need not be true, just highly probable. Note that by making things like "truth" an epistemic value, we are also claiming that truth is not in the world in some sense of "in". Rather, truth is a property of claims about the world, as is "highly probable". And we are the ones making claims about the world. Often we make incorrect claims about the world for a variety of reasons such as not having enough evidence, or not calibrating our instruments correctly. The conclusion is that if scientific inquiry, often deemed by many as our best knowledge-generating activity, is value laden all the way down, then what does that say about other forms of decision-making?

## 6.12 Decision-Making

Choosing A over B, no matter what the circumstance, will always involve values. If asked to explain your choice of A, you will appeal to a value. To see this, consider the structure of decision-making.[8] Basically we approach decision-making already equipped with items: background knowledge, values (motivators to act in certain ways) and goals. We ask ourselves which choice among the options will best help us secure our goals. We use our background knowledge and our values to evaluate our options. That is, I know roughly how these two machines work – let us say that they will both do the job, but one costs double the other. Our goal is to make money, so picking the cheaper machine appears more conducive to achieving our goal.

Now it is one thing to make a decision and another to make a good decision, or, as people are irrationally fond of saying, a rational decision. There are numerous theories of rationality and short of engaging in a detailed critique of each, I will lay out my own account and go with that. I approach the problem of defining "rational" by considering what it is to be irrational. A key example of being irrational is not learning from experience. If you make a decision to do X and doing X fails to get you the desired result, it is irrational to simply do X again and again, hoping that one of these times it will work. The rational person, having failed will examine her assumptions to try to figure out why what seemed to be a reasonable choice failed. What the person is engaged in at this point is a feedback loop. You take what you have learned from doing X and reevaluate your decision to do X in light of that new knowledge. Let's say that what you learn is that while it is true that the machine you chose was cheaper than the other, it was also not as sturdy and could not hold up to the burden placed on it. You also realize that having chosen the cheaper machine out of greed, you now ended up losing money because you now have to buy the costlier as well. At this point you begin to reconsider the value of making money no matter what and decide that in the future successfully completing the job is more important than failing.

[8] I have elaborated this view in my 2000/2006.

So the structure of decision-making is this: start with knowledge,[9] values and goals – you choose an option to act on – you act – you evaluate the results of your action. If everything came out ok – nothing you know, value or have as a goal needs to be changed. But if you failed to achieve your desired result, something needs to be reconsidered. So learning from experience is taking the results of failure and using them as a basis for reevaluating your knowledge and your values and your goals – not all at once, but little by little until you get things sorted out.

As you can see, values are built into decision-making. I call this view the Commonsense Principle of Rationality or CPR. It is based on how we do in fact reason when we are reasoning successfully. It also has a normative component in that I am also suggesting that this is how you ought to reason.

If I am right, every decision is value-laden in that values constitute a significant component of every decision. That being the case, if we turn to the problem of developing a technological artifact, we can see how this view leads us to conclude that we can't really talk about values being embedded in an artifact in any meaningful way. Here's why.

## 6.13 A Plethora of Values

When we talk about the process of making technological artifacts, the road from the glimmer of an idea to successful production is long and complicated. Consider an engineering context. Someone in your company sees a call for proposals to build a whatsit. Essentially the description of the item is in terms of what it is supposed to do. We need a whatsit to do blah, blah, blah. Specifications are as follows: it must not weigh more than x; it must fit into the following sized compartment; it must come in at or below (some monetary figure), etc. Management decides to turn in a bid. It calls the design team together and tells them to come up with a design that meets function and specs. The design team gets to work. This entails a lot of brain storming – a lot of proposals – a lot of arguments about whether this proposal or that will do the job. At each stage in the design process various options are considered by numerous members of the team and many decisions are made and remade. I propose that at no point can you point to the value that went into the object or that the object embodies because there are many and design is only part of the process leading to the eventual design (see Buccarrelli 1994). Everyone involved in deciding something about the design is making decisions that necessarily involve values, as we saw above. Now, if we were to concede that it is possible for values to be embedded in artifacts, there will be lots of them, so many in fact that it is unlikely they can be differentiated. In short, even conceding that artifacts embody values gives us little to work with. One should probably shrug and say "so what?" If we move to the next stage it gets more

[9] It is also important to note that the knowledge you have at the start is not just a set of abstract propositions, it also illuminates the context in which you are operating. You know, for instance, that you are here, not there, that you have the following items to contend with, etc. Since the mark of knowledge is successful action, it is also the case that since actions are contextualized, so is knowledge.

complicated because there are decisions made as to whether to submit the final design, decisions to accept the design (usually with modifications), then we move to production and the various individuals involved in those decisions, et., then marketing. Robert Moses may have desired to keep inner city people from the beaches on Long Island, but the values embedded in the LIE, if values are embedded in things come from many people and who can tell which ones are Moses’?

## 6.14 Conclusions

So, where are we? VNT claims that values are not embedded in technological artifacts. That is a very strong claim. A weaker version (2) claims that even if we could make sense of the idea that technological artifacts embody human values, there are so many that would be involved the claim says nothing significant. There is also a third version (3) that says we don’t know whether or not values are embedded in technological artifacts because we don’t know what values look like. While I have explored some aspects of (3), I have more seriously been elaborating a defense of version (2). That is, since artifacts are the results of human decisions and since human decisions are a function of human values, understood as motivators to achieve a certain preferred state of affairs, and since many people are involved in the creation of technological artifacts, it adds nothing to the discussion to say values are embedded in artifacts.

One of the interesting features of human discourse is its capacity to mislead. My final diagnosis of the slogan “Guns don’t kill, people kill” is so hard to agree or disagree with is that it trades on our desire for simple answers. While it is true that guns don’t kill all by themselves, in this context it is also true that people rarely kill all by themselves. But the slogan seems to suggest that we are confronted with an either/or situation. Either people kill or guns kill. But putting the situation to us in that way only forces a false choice. Perhaps a better way to put it is “Guns don’t kill, people kill using guns, knives, their hands, garrotes, automobiles, fighter planes, poison, voodoo dolls, etc.” The culprit is people.

## References

Baird, D. (2004). *Thing knowledge*. Berkeley: University of California Press.
Buccarrelli, L. (1994). *Designing engineers*. Cambridge: MIT Press.
Hume, D. (1738/1978). *A treatise of human nature*. London: Oxford.
Lewis, C. I. (1946). *An analysis of knowledge and valuation*. La Salle: The Open Court Publishing Company.
Pitt, J. C. (2000/2006). *Thinking about technology*. New York: Seven Bridges Press. http://phil.vt.edu/Pitt/jpitt.html
Pitt, J. C. (2005). Hume and Peirce on belief, or, why belief should not be an epistemic category. *Transactions of the Sharles S. Peirce Society, XLI*(2), 343–354.
Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, 1–6.
Woolgar, S., & Cooper, G. (1999, June). Do artefacts have ambivalence? Moses’ bridges, Winner’s bridges, and other urban legends in ST&S. *Soc Stud Sci, 29*(3): 450–457.

# Chapter 7
# Can Technology Embody Values?

**Ibo van de Poel and Peter Kroes**

**Abstract** Under the banner of Value Sensitive Design (VSD) various proposals have been put forward in recent times to integrate moral values in technology through design. These proposals suppose that technology, more in particular technical artifacts, can embody values. In this contribution, we investigate whether this idea holds water. To do so, we examine the neutrality thesis about technology, that is, the thesis that technology is neutral with regard to moral values. This thesis may be interpreted in various ways depending on the kind of values involved. We introduce two distinctions with regard to values: (1) final value (value for its own sake) versus instrumental value, and (2) intrinsic value (value on its own) versus relational or extrinsic value. This leads to four different kinds of values to which the neutrality thesis may refer. We argue that the most interesting version of the neutrality thesis refers to extrinsic final values. We provide a number of counterexamples to this version of the neutrality thesis, and, on the basis of these examples, we suggest a general account of when a technology may be said to embody values. Applying our results to VSD, we introduce three different values involved in a design process, namely, intended value (the value intended by the designers) embodied value (the value designed into the artifact) and realized value (the value that is realized in actual use) and we discuss how we can verify what values are embodied in a designed technical artifact.

I. van de Poel (✉) • P. Kroes
School of Technology, Policy and Management, Philosophy Section, Delft University of Technology, Delft, The Netherlands
e-mail: i.r.vandepoel@tudelft.nl; p.a.kroes@tudelft.nl

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_7,

## 7.1 Introduction

In recent times various authors have argued for taking into account considerations about moral values in the engineering design process by what they call Value Sensitive Design (VSD). This is an approach that aims at integrating values of ethical importance in a systematic way into the designs of technical artifacts (Friedman 1996; Friedman and Kahn 2003; Friedman et al. 2006). The approach has been applied to a number of design projects especially in information and communication technologies (ICT) but the basic idea of the approach is more generally applicable.

A central tenet of VSD is that we can somehow design moral values or values in general into technical artifacts, so that they can embody values (cf. Flanagan et al. 2008). This assumption is, however, not uncontroversial. Our main aim in this contribution is to critically assess the idea that technical artifacts may embody values, in particular moral values. We will do so by contrasting this assumption with the neutrality thesis of technology. One of the most powerful expressions of the neutrality thesis is contained in the slogan of the American National Rifle Association: "Guns don't kill people, people kill people". This statement is not intended to deny that guns can be used for morally good or bad purposes; they can. Rather it holds that it is this use, and not the technology itself, that is morally good or bad and thus has moral value. In its most general form this neutrality thesis with respect to technology can be expressed as follows:

(N) Technology is morally neutral.

The meaning of N depends, of course, on the meaning of the notion of technology and what it means to be morally neutral. There are various ways in which we may interpret the notion of technology (see for example Mitcham 1994). Here we will take technology to be a collection of technical artifacts – we will have more to say on the notion of technical artifact below. Furthermore we will assume that something is morally neutral if and only if it does not embody moral values. With regard to VSD, the interesting question is not whether *all* technical artifacts are morally value-laden (or *all* are morally value-free) but rather whether it is possible to make some technical artifacts morally value-laden by consciously designing them that way. We therefore propose to reformulate the neutrality thesis N as follows:

(N1) Technical artifacts cannot embody moral values.

If N1 is true, it is not possible to design values into technical artifacts and therefore the basic assumption underlying VSD is ill-founded. The most obvious defense of N1, it seems, starts from the assumption that technical artifacts are mere instruments.[1] As mere instruments, they can be used for morally good or bad ends, but technical artifacts themselves, independent of these ends, are value neutral; they do

[1] It is hard to find explicit defenses of the neutrality thesis in the literature, but see Pitt (2000) and Pitt's contribution to this volume.

not by themselves embody moral values. So, technical artifacts may have instrumental value and this instrumental value may be exploited in realizing ends that may be assessed as morally good or bad. In terms of the use plan interpretation of technical artifacts developed by Houkes and Vermaas (2010) this may be expressed by saying that only because of the goal of the use plan in which a technical artifact is embedded, technical artifacts may have moral significance. So, it is their (social) context of use that determines whether technical artifacts have moral values.

In this defense there are two issues at stake. One issue is whether technical artifacts can only embody instrumental value or also what we will call below final value, i.e. value for its own sake. The other issue is whether technical artifacts by themselves can have certain values (intrinsic value), or whether they can have values only in relation to something else (relational value). To understand the neutrality thesis, and to affirm or rebut it, we need to disentangle both aspects. We therefore start this article with a rather long philosophical detour aimed at better understanding the notion of value. This detour will enable us to formulate the neutrality thesis more precisely. We will then provide a number of examples that rebut the neutrality thesis. Having argued that technical artifacts may embody a particular kind of values, we return in the final part briefly to VSD and analyze how values may be embodied in technical artifacts by relating them to their designed features.

## 7.2 Moore on Intrinsic Value

We start our philosophical detour about values with G.E. Moore's characterization of intrinsic value. The reason is not that we subscribe to Moore's characterization of intrinsic value. Rather, we believe, like various other contemporary philosophers, that Moore's conception of intrinsic value is somewhat confusing, for reasons we will explain below. Nevertheless, Moore's characterization is interesting because it has been quite influential in philosophy and because it appears to touch upon both aspects in the debate about the neutrality thesis we alluded to above, i.e. instrumental versus final value and intrinsic versus relational or extrinsic value.

In the beginning of the twentieth century, G.E. Moore has offered the following account of intrinsic value (Moore 1903, 1912, 1922).[2] Moore believed that 'goodness' (the term he used for what we call value) is an unanalyzable property; in particular it cannot be defined or analyzed in terms of natural or descriptive properties. At the same time, Moore believed that goodness was objective and did not depend on people's desires or appreciations. This brought him to the notion of intrinsic value, as value that is intrinsic to the valuable object.[3] For Moore intrinsic

[2] For a discussion of different notions of intrinsic value that have been distinguished by philosophers, see Feldman (2005).

[3] There is a debate in the philosophical literature about what kind of entities can bear value; some believe that only states-of-affairs can be bearers of values, others, like Moore, also include, for example, objects. We will not enter into this debate here, but we will assume that objects can be

value was not a property of an object, at least not a descriptive or natural property because he is a non-naturalist about goodness. He nevertheless seems to believe that intrinsic value depends on the intrinsic (natural) properties of an object. One possible way of expressing this idea is by saying that intrinsic value supervenes on intrinsic natural properties but cannot be analyzed in or reduced to these natural properties. According to Moore, then, a value that is intrinsic to an object remains the same whatever its relation to other things: "A kind of value is intrinsic if and only if, when anything possesses it, that same thing or anything exactly like it would necessarily or must always, under all circumstances, possess it in exactly the same degree" (Moore 1922: 265). For Moore, intrinsic value is thus by definition not extrinsic or relational.

Although the emphasis in Moore's account lies on what it means for an object to have intrinsic value, he also assumed that only so-called final values can be intrinsic values. The reason for this assumption seems rather straightforward: things with instrumental values derive their value from them being instruments for attaining something else that is valuable (be it for its own sake or not). So instrumentally valuable objects by definition derive their value from something outside the object, and therefore the value of these objects is not intrinsic to those objects, but relational.

## 7.3 Various Forms of Value

For Moore then the notion of intrinsic value combines two aspects: (1) value that is intrinsic to an object, i.e. value that only depends on an object's intrinsic properties and (2) final value, i.e. value for its own sake. Especially Christine Korsgaard's 1983 article "Two Distinctions in Goodness" has drawn attention to the distinction between these two aspects (see also Kagan 2005; Rabinowicz and Rønnow-Rasmussen 2005). Korsgaard follows Moore in saying that objects that are valuable due to their intrinsic properties are unconditionally good (Korsgaard 1983). Their goodness does not depend on their relation to other objects or to people; otherwise their value would not be intrinsic to the object. However, according to Korsgaard, some things may be good for their own sake, even if they are not unconditionally good. An example is human happiness understood in a Kantian way. According to Kant, human happiness is good for its own sake; happiness is not an instrumental value but a final value. Nevertheless, according to Kant, happiness is only conditionally good; it is only good insofar as brought about by the good will, i.e. out of respect for the moral law.

Taking into account the distinction that Korsgaard refers to, we propose to classify the values of objects in two independent ways. The first relates to whether values are

bearers of value (otherwise the neutrality thesis seems obviously true). In the text we will refer to objects as bearers of value but we do not want to imply that only objects, and not, for example, states-of-affairs or persons, can be bearers of value.

**Table 7.1** Types of value

| | Intrinsic value (Non-relational) | Extrinsic value (Relational) |
|---|---|---|
| Final value (For its own sake) | Intrinsic final value | Extrinsic final value |
| Instrumental value (Not for its own sake) | Intrinsic instrumental value | Extrinsic instrumental value |

relational or not. Values that are not relational will be called 'intrinsic values' because these values depend only on intrinsic properties. Otherwise, values are called 'extrinsic'. The second way relates to whether the values of objects are values for their own sake or not. Values for their own sake will be referred to as 'final values'; otherwise values will be called 'instrumental values'. Doing so, we end up with the following four possible combinations of values (see Table 7.1).

Whether or not it is possible to make sense of all four combinations of values is an issue that falls largely outside the scope of this paper. With regard to the neutrality thesis N1 and VSD the interesting question now is what notion of value might be at stake in these claims. We will argue that N1 can best be understood in terms of final values, and that technical artifacts may embody (moral) extrinsic final values, which means that N1 does not hold. But rather than getting ahead of our argument, we will first argue why we believe that N1 should not be interpreted as referring to instrumental values. One rather straightforward reason would be to note that many defenders of N1 do not want to deny that technical artifacts have instrumental value. We think that we should, however, dig a bit deeper, for there appears to be a more fundamental reason to assume that N1 is not about instrumental values: instrumental values may well be not values at all.

## 7.4 Are Instrumental Values Real Values?

Several philosophers have suggested that instrumental value is not a value at all. Some of them seem to assume that it is obvious that instrumental values are not real values (Moore 1903; Ross 1930; Dancy 2005b). Others suggest that the idea of instrumental value being a value is based on a linguistic or terminological confusion. Instrumental value refers not to being a value but rather means something like "being a means to" (Rønnow-Rasmussen 2002). Below we will try to provide an argument why instrumental values are not 'real' values. At the bottom of this argument lies an assumption about a relation between values and reasons to which many modern philosophers seem to ascribe. We start with setting out this relation and then apply it to instrumental values.

Both values and reasons belong to the normative domain; they belong, however, to different parts of the normative domain. Values come in different kinds, such as epistemic value (truth), aesthetic value (beauty), pragmatic value (efficacy/efficiency) or moral value (moral goodness). What these values have in common is that

they are varieties of goodness (von Wright 1963). It is on the basis of values that we evaluate certain objects or state-of-affairs as good or bad and beautiful or ugly. Values, therefore, have their home in the evaluative part of the normative domain. Reasons, however, belong to the deontic part. Reasons relate to what to do, believe or aim for. Reasons are considerations that count in favor of or against doing, believing or aiming for something. Reasons are to be distinguished from 'oughts' or obligations, which also belong to the deontic domain. If one has reason to do something one is usually not obliged to do it (although different authors sometimes use somewhat different terminology here). Often there are both reasons for and reasons against doing something and an 'ought' is then believed to result from the totality of relevant reasons, although the totality of reasons can also be inconclusive or can merely allow to do something without there being an obligation to do it.

There is no agreement in the philosophical literature on how values and reasons are related. One category of theories, often called 'consequentialism', holds that we have reason to do what has or brings about value, that we should increase the amount of value in the world or even should maximize it. Such theories thus believe that values precede reasons: they are what give us reasons. One need, however, not be a consequentialist in the above sense, to maintain that values are metaphysically prior to reasons. Jospeh Raz, for example, holds that values give us reasons to engage with those values in appropriate ways (Raz 1999). What appropriate is may, however, depend on the value and the situation: some values are to be promoted or maximized (as consequentialists hold), other are to be admired, cherished or enjoyed.

Other theories hold that reasons are metaphysically prior to values. Elisabeth Anderson, for example, defends what she calls an expressive theory of rational choice (Anderson 1993). According to her statements like 'x is good' or 'x is valuable' can be reduced to 'it is rational to adopt a certain favorable attitude towards x.'[4] The reasons we have to adopt certain attitudes to things or state-of-affairs ground the value of those state-of-affairs or things. A somewhat different account is offered by Scanlon, who argues that "being good, or valuable, is not a property that itself provides a reason to respond to a thing in a certain way. Rather, to be good or valuable is to have other properties that constitute a reason" (Scanlon 1998: 97).

We will not take a position in the theoretical debate about the relation between reasons and values here. It is, however, worth noting that all positions we briefly mentioned seem to suppose a certain correspondence between values and reasons of the following kind:

(V) If x is valuable (in a certain respect) then one has reasons (of a certain kind) for a positive response (a pro-attitude or a pro-behavior) towards x.

This statement is intended to be neutral with respect to the question whether values ground reasons or reasons ground values or that neither can be reduced to the other. As Dancy (Dancy 2005b) notes, whatever position one takes in this debate something like V seems to be true. The notion of positive response in V is meant to

[4] We might also have a reason for a negative rather than a positive response. This would then be associated with disvalue rather than value.

capture a range of pro-attitudes and pro-behaviors like desiring, promoting, caring for, admiring, enjoying, loving et cetera. As suggested above what positive response is adequate depends on the kind of reasons or values (and the situation).

What makes V interesting for our current purpose is that it may have a certain pragmatic or epistemological relevance for tracking or recognizing values. If we want to know whether a certain x is valuable (bears or embodies a value), we need to check whether there are reasons for a positive response towards x. If such reasons are absent, x has no value. Of course, if there are such reasons then V does not imply that x is valuable. The so-called 'wrong kind of reasons' problem (cf. Schroeder 2009) illustrates that one should be careful not to reverse the implication in V. For example, if I promise someone to give him an object x tomorrow this gives me a reason to protect x now (for example against theft) and protecting expresses a positive response to the object x. This reason, however, is based on my promise and in no way related to the object x itself (apart from it being the object of my promise). It is therefore the wrong kind of reason to track the value of x. Therefore not all reasons for positive responses towards x track or indicate that x is valuable, at least we need to make sure that the reasons relate to x itself and not to something else.

With respect to instrumental value, the crucial question is whether the instrumental value of an object provides reasons for a certain positive response to that object. For example, is the instrumental value of a knife for cutting a reason to use it for cutting?[5] Not as such, but it may be if I desire to cut something; then the instrumental value of the knife may be a reason to use it for cutting. However, as several philosophers have pointed out, the fact that I desire to do something is as such not a reason to do it (Raz 1986; Scanlon 1998; Dancy 2002). This is not to deny that I might have a reason to do what I desire but this reason is not grounded in the desire but in something else; the fact that I have the desire as such does not add anything to my reason. So desiring to cut cannot provide the right reasons for cutting nor for using the knife for cutting. From this it follows that instrumental value cannot be associated, at least not always, with reasons. This may be taken as a strong argument why instrumental value is not a real value. In the appendix we discuss in more detail why the instrumental value of technical artifacts cannot be associated with reasons, or at least not with the right kind of reasons.

## 7.5 A Reformulation of the Neutrality Thesis in Terms of Extrinsic Final Value

We have identified four possible forms of value (Table 7.1) and we have argued that there are good reasons to doubt that the instrumental forms of values are real values. This leaves us with two forms of values, intrinsic final values and extrinsic final

[5] We take 'using' here to be a positive response. For further discussion, see the Appendix.

values, to which the neutrality thesis might refer. Let us first look at the interpretation of the neutrality thesis in terms of intrinsic final values:

(N2′) Technical artifacts cannot embody moral intrinsic final values.

The problem with N2′ is that it appears hard if not impossible to deny. The idea that a technical artifact has a form of value that remains the same independent of its relation to anything else, in particular of its design context or its context of use is very implausible. A serious problem with regard to N2′ is that it is not clear at all what kind of value could undermine N2′. If no conceivable value can be intrinsic to technology, then N2′ runs the risk of being true by definition. This means that N2′ as our construal of the idea that technology is value-neutral is more or less a truism.

The foregoing is related to a conceptual point about technical artifacts. In a nutshell this point is the following. Roughly, technical artifacts may be characterized as physical objects with a practical function. Typically, the physical object is a human made physical construction. But not any physical construction made by humans is a technical artifact; for that it is necessary that that physical construction is to be used for doing something, that is, that it has a technical function. Neither is a function without a physical construction that realizes that function a technical artifact. Both the physical structure and function are constitutive for being a technical artifact. This means that a technical artifact has a dual nature: it is a hybrid object with physical and functional features (see Kroes and Meijers 2006; Kroes 2010). Now, the physical features are intrinsic features of a technical artifact, but that is not true for its functional features. On the one hand, its functional features are related to its intrinsic physical features, because the physical structure has to realize the function. But, on the other hand, the functional features are related to human intentions or practices of intentional human action. It is only in relation to human intentions that technical artifacts have functions. More in particular we assume in the following that the intentions of designers, and not those of users, are constitutive for an object to be an instance of a particular technical artifact kind (for more details, see Kroes 2012). However, irrespective of whether the intentions of designers or users play this role, being a technical artifact involves intrinsic as well as relational properties.

According to the dual nature account, technical artifacts cannot be conceptualized or characterized fully in terms of their intrinsic physical properties alone. What distinguishes a technical artifact from a mere physical object are some of its relational or extrinsic properties. Such extrinsic properties, however, cannot be the ground for any intrinsic final value of a technical artifact. This means that, in so far a technical artifact has intrinsic final value it must have this value in virtue of its physical properties, that is, in virtue of being a physical object. So N2′ is not so much a statement about technical artifacts as well a statement about physical objects. Since it is generally assumed that physical objects, qua physical objects, have no intrinsic value, N2′ may be true, but it is not a very interesting thesis about technology or technical artifacts because it disregards those (extrinsic) features of physical objects that make them technical artifacts. A similar conclusion may be drawn on the basis of the use plan approach to technical artifacts. According to

Houkes and Vermaas (2010) what makes a physical object into a technical artifact is the fact that it is embedded in a use plan; without a use plan, no technical artifact. This feature of a technical artifact, of being a physical object embedded in a use plan, however, is a relational or extrinsic feature, not an intrinsic one; it relates technical artifacts to human beings. So, again, since any intrinsic final value of a technical artifact will have to be grounded in its intrinsic features, it follows that in so far a technical artifact would have any intrinsic final value, it would have so in virtue of being a physical object.

Let us shift our attention from intrinsic final values to extrinsic final ones. Then we end up with the following version of the neutrality thesis

(N2″) Technical artifacts cannot embody moral extrinsic final values.

A first thing to note is that the notion of extrinsic final value is not uncontroversial. Nevertheless, various philosophers have argued for the existence of extrinsic final values (Korsgaard 1983; Kagan 2005; Rabinowicz and Rønnow-Rasmussen 2005). We will not consider their arguments in detail, but cite two kinds of examples that make their argument plausible. One kind of example concerns cases in which something has final value, or at least more final value, than it would otherwise have because it is rare.[6] A rare stamp has more value than a regular stamp. The last remaining vase from a certain time period has final value not so much because of its intrinsic properties but because it is the only exemplar left. Given that rarity is a relational property rather than an intrinsic property these examples suggest that something like extrinsic final value is possible. Another kind of example concerns objects that have value because they belonged to a particular person, for instance, my mother's wedding ring, which again is a relational rather than an intrinsic property.

These examples can easily be extended to technical artifacts. A rare car from the 1920s may have final value because of its rarity. Similarly, the guillotine which with Louis XVI was killed may have historical final value. These kinds of examples raise, however, another worry. They are not the right kind of examples to reject the neutrality thesis because they do not refer to the specific technical or designed features of the technical artifacts involved. It appears that we somehow must restrict the extrinsic or relational properties on which the final value of a technical artifact may supervene to get an interesting version of N2″. We propose to do so by adopting, and slightly (but significantly) revising, a proposal that Dancy has done to distinguish between what he calls the resultance base and the supervenience base of a value. Dancy introduces this distinction because he wants to allow for the fact that a feature "may have one value in one context and a different or [even] opposite value in another" (Dancy 2005a: 333). At the same time Dancy wants to retain something of Moore's original idea that value supervenes on intrinsic properties. He therefore distinguishes "between those features from which some value results (the good-making features, as we might put it), and other features whose presence or absence would

[6] Keep in mind that according to Moore two similar objects should not just have both intrinsic value but also exactly the same amount of intrinsic value.

have made a difference." The first features or properties form the resultance base: they generate the value. The second type of features are the supervenience base and "can make a difference to the ability of the intrinsic properties to generate the value that they do" (Dancy 2005a: 334).

Dancy appears to equate the resultance base with intrinsic properties. This proposal will not do for our purpose because, as we have seen above, some of the defining properties of technical artifacts are extrinsic in nature. Nevertheless the notion of resultance base can be used if we adapt it to refer to those properties that define the technical artifact, excluding from the resultance base those relational properties that a technical artifact has by virtue of its specific context of use. These specific contextual properties still might be considered part of the supervenience base and they may influence the ability of the properties in the resultance base to actually generate the value they potentially do. In this way, we can allow for the context to make a difference for the value that is actually realized while at the same time we can maintain the claim that a technical artifact has a value that is generated by the technical artifact itself rather than its context of use. The latter value may be a value that a technical artifact has for its own sake, that is, may be a final value. Nevertheless, such a final value will be relational or extrinsic because it is grounded in a resultance base that is partly relational.

Our conception of the resultance base may leave open the possibility of a technical artifact having extrinsic final value in general, but we still have to define the restrictions to be put on the resultance base in order to arrive at an interesting version of N2″. One possibility would be to focus on those properties that are (minimally) necessary to call something a technical artifact. That may be the right choice if one wants to know what values may be embodied by technology in general or by technology as the class of all technical artifacts. Our purpose here is somewhat different: we are interested in whether it is possible to embody specific values in technical artifacts through design (VSD). We will therefore interpret the resultance base of a particular technical artifact as those properties that are designed into that object. If these designed properties can indeed generate value, we have reason to suppose that we can embody value in technical artifacts by design and that VSD is possible. This brings us to the following reformulation of the neutrality thesis:

(N3) The designed properties of technical artifacts cannot form the resultance base of moral extrinsic final values.

Below, we will argue against N3. Before we do so, it is worthwhile to consider what denying N3 would and would not imply. First, the denial of N3 does not entail that all technical artifacts embody extrinsic final value. Rather it implies that technical artifacts can embody such values and that this embodiment can be achieved through design. Second, the denial of N3 does not imply that technical artifacts embodying extrinsic final values will always realize these values in actual practice. According to the adapted version of Dancy's distinction this is dependent on the entire supervenience base that includes the extrinsic properties related to the context of use as well. So, denying N3 implies that the potential to generate certain specific values can be embodied in certain technical artifacts.

## 7.6 Rebutting the Neutrality Thesis: Some Examples

We will now rebut the neutrality thesis N3 through a number of examples. Before we do so, some clarifications are in order. First, as noted above, N3 and its denial are claims about the resultance base and not about what values are realized in practice. To deal with this, we propose to make the following terminological distinction. We will use the notion *realized value* as the value that is realized by a technical artifact in a practical context; the realized value is dependent on the entire supervenience base as argued above. We will use the notion of *embodied value* as the value that results from the resultance base; an embodied value is not necessarily realized in an actual context. Embodied value may be understood as the potential to realize a value in an appropriate context. We have more to say on the distinction between embodied value and realized value in the final section, but for the moment this basic distinction suffices.

Second, we will take the designed features of a technical artifact to be intentionally designed features (unless stated otherwise). This might seem obvious because design is an intentional activity. However, even if design is intentionally directed at creating technical artifacts with certain features, it does not follow that all the designed properties are necessarily intended properties. Cars, for example, pollute the environment and this may be considered a feature that results from the design of cars, but this feature is not intended, at least not in the common sense notion of intending. We do not want to enter into a philosophical discussion on the notion of intention here, but simply postulate that below we will be focusing on the intentionally designed properties of technical artifacts. Even if there are also unintentionally designed properties, this does not pose a problem for our undertaking. We are looking for examples that rebut N3. Since the intentionally designed properties of a technical artifact are obviously a subset of its designed properties, examples of intentionally designed features are ipso facto examples of designed features and, therefore, they are relevant for rebutting N3.

Third, we will make reference to functional features or functions of technical artifacts. We are aware that various function theories interpret functions in different ways, ranging from intended physical capacities through intended behavior to intended effects and purposes (see Houkes and Vermaas 2010 and Van Eck 2011). For our purposes it will not be necessary to commit ourselves to any particular function theory.[7] Note moreover that functions are usually associated with instrumental values, since they are interpreted in terms of means-ends relations. Below, however, we will associate functions also with final values.

With these clarifications in place, we can now turn to our task of presenting a number of examples that rebut N3. The first category of examples we will provide are examples in which the embodied extrinsic final value of a technical artifact

[7] We do, however, exclude function theories that identify functions with physical capacities, for those theories would make functions intrinsic properties of technical artifacts. Function theories that identify functions with intended capacities are, however, not excluded, since intended capacities are not intrinsic properties.

coincides with, or is hardly distinguishable from, its function. These examples are based on the assumption that it is uncontroversial that the function of a technical artifact results from its designed features. Now, if we can show that in some cases the extrinsic final value of a technical artifact is indistinguishable from its function, we have shown, contrary to N3, that a technical artifact's designed features may form the resultance base for extrinsic final value, which means that a technical artifact can embody such values.

The first example concerns sea dikes. The technical function of a sea dike is to prevent the hinterland from flooding (e.g. Herbich 1999: 3.4). Protecting the hinterland from flooding is instrumental to a moral value like the safety of the inhabitants of the hinterland, which we consider to be a final value. The point is not that sea dikes can be used to achieve safety but that achieving safety is part of their *function*. This is witnessed by the fact that design requirements, and in fact legal norms, and design approaches for dikes are based on the value of safety (Snippen et al. 2005). Dikes are thus *designed for safety*. This is different from, for example, a knife. The function of a knife is cutting; cutting of, for example, bread may be instrumental to a final value like health or survival or human-well-being. However, the attainment of such final values neither is part of the function of knifes nor have normal knifes been designed to achieve such final values. Whereas in the case of the knife, the function of the artifact and the final values that can be achieved by realizing the function are clearly separated this is not the case in the sea dike example. The instrumental function of sea dikes (protection from flooding) can hardly be distinguished from the final value for which they are designed (safety with regard to flooding). After all, the technical function of a dike may be described as providing safety with regard to flooding. If such expressions make sense, then it follows immediately that technical artifacts, as objects with a function, may embody extrinsic final values, since functions are extrinsic features of technical artifacts.

A second example is the speed bump. The function of speed bumps is to slow down cars in, for example, living areas and this is conducive to traffic safety, which again we assume to be a final value.[8] Similar to the dike case, being conducive to traffic safety is not just an instrumental feature that speed bumps happen to have but it is a purposively designed feature, it is what speed bumps are designed and used for. Moreover, like the sea dike example, the function of the speed bump (slowing down cars) is hard to distinguish from the final value to which it is instrumental (traffic safety). So, also speed bumps may be said to embody an extrinsic final value, namely that of traffic safety. That they indeed embody this value is also confirmed by the fact that we appear to have certain reasons to positively respond to speed bumps given the fact that they are designed for traffic safety. Suppose that someone feels inclined to speed over speed bumps because he likes a bumpy ride or he likes the kick of dangerous driving. Such a person does not seem to respond properly to speed bumps because they are designed (intended) to let people slow down and to

[8] See e.g. http://www.portlandonline.com/Transportation/index.cfm?a=83939&c=38764#function. Accessed December, 14 2009.

increase traffic safety. In other words, speeds bumps give us reasons to slow down not just because it is inconvenient to drive fast over a speed bump but primarily because they have the function of traffic safety.

Someone might object that we have a reason to slow down in living areas anyway, whether there are speed bumps or not. This is true, but our point is that the speed bump and its intimate connection to traffic safety give an *additional* reason to *respond* to the speed bump in a specific way, i.e. by slowing down. This response is the expression of a pro-attitude because it respects the function/value of the speed bump and it therefore fits thesis V. Another objection might be that whether this is indeed the proper response will also depend on the use context. Suppose that a speed bump is part of a racing track to add an element of skillful driving to a racing competition. In that case, slowing down does not seem the proper response, but it is rather something like skillfully driving as fast as possible over the speed bump. We agree that in those circumstances, the value of, and the proper response to the speed bump are different from the normal circumstances. This difference, however, can be understood in terms of the difference between resultance base versus supervenience base introduced earlier. The claim is, then, that the value of traffic safety results from the resultance base, i.e. the designed features, of the speed bump while the supervenience base, that determines whether this value is indeed realized in practice, also depends on the context of use.[9]

What is crucial to these examples is that the final values involved are part of the function of a technical artifact. It does make sense to say that the function of dikes is the safety of the hinterland and of the people living there or that the function of speeds bumps is traffic safety. There are, however, also cases in which the function of an artifact may be instrumental to a final value but in which the final value is itself not part of the function. Take for example a hygrometer. The function of a hygrometer is to measure humidity. Measurements of humidity can be used, for example, to protect valuable paintings in museums. Protecting valuable painting is a final value (we suppose). It would, however, not make sense to claim that the function of a hygrometer is to protect valuable paintings. (Maybe the function of 'museum hygrometers', if such technical artifacts would exist, may be said to protect valuable paintings). Moreover, the use of a hygrometer for another purpose than protecting paintings seems in general not improper while using speed bumps for reckless driving seems an improper response in normal circumstances.[10] So unlike sea dikes and speed bumps, hygrometers do not embody final values.

[9] In the final section, we will discuss in more detail how one can determine whether a certain value indeed results from the resultance base even if it is not always realized in practice.

[10] It might be inappropriate not to use a hygrometer for protecting valuable paintings in certain circumstances, but in such cases it is an inappropriate response to the value of paintings rather than to the value of the hygrometer.

## 7.7 Side-Effects

We now turn to a second category of examples. In these examples the final value is not part of an artifact's function, but it nevertheless results from its designed features. A first example in this category are the low overpasses at the Long Island parkways designed by city builder Robert Moses, as discussed by Langdon Winner (1980). According to Winner, Moses intentionally designed these overpasses extraordinary low for racist motives. The low overpasses would make it impossible to reach the beaches by public transport because buses could not pass below them. So, only people who could afford a car – and in Moses' days these were generally not Afro-American people – could easily access the beaches.

Winner's interpretation of this case is contested (e.g. Joerges 1999). It has been questioned whether Moses really made the bridges low for racist motives or that he maybe did so on the basis of other considerations. It is also not clear whether it was really impossible to reach the beaches by public transport as a result of the low overpasses. For the sake of the argument, we will nevertheless accept Winner's version of the story; after all it seems conceivable that some city builder designs low overpasses for the reason and to the effect that Winner ascribes to Robert Moses.

Now, the question is whether it makes sense to say that the low overpasses at Long Island embody the value, or rather disvalue, of racism.[11] Obviously, it is not the technical function of the low overpasses to prevent Afro-Americans to reach the beaches, or even to make impossible public transport over the Long Island parkways. These are rather side-effects.[12] In general the occurrence of side-effects seems not enough to ascribe the associated value or disvalue, in this case racism, to the technological artifact that causes the side-effects. One reason why such ascriptions seem dubious is that the side-effects may arise from the specific way an artifact is used or from its employment in an unusual context. This case is, however, not just an example of side effects but it is an example of intended side-effects (on Winner's reading at least).

We believe that it makes sense to say that the overpasses embody the disvalue of racism. One reason to think so is that the overpasses are intentionally *designed for* racism. This intentional history gives the overpasses a certain meaning or symbolic value, which corresponds with reasons to disapprove of them. Similarly, the fact that the gas chambers in German concentration camps during the Second World War were designed to contribute to the extinction of the Jews gives us reason to abhor those gas chambers. It might be objected that our disapproval in such cases concerns the intentions of the designers rather than the technical artifact itself. Surely, we also have reasons to disapprove the intentions of the designers,

[11] The reason why we analyze this case in terms of disvalue and not of value is that V is formulated in terms of pro-attitudes and racism does not correspond with pro-attitudes but rather with contra-attitudes (at least for most people we hope) which may be associated with disvalue (or negative value).

[12] The function of an overpass is something like the crossing of one road over another. Making overpasses extraordinary low does not change this (basic) function.

but we also believe that there might be independent and additional reasons to disapprove the technical artifact itself, at least in those cases that the artifact has the potential to realize the intended disvalue as a result of its designed properties. If the overpasses in Winner's example lacked the capacity to prevent buses (and so Afro-Americans) to go to the beach or if the German gas chambers lacked the capacity to kill Jews, we might still disapprove the intentions of the designers but not the artifact itself. The importance of this condition is even clearer in cases of a positive value. We may admire or cherish pace makers because they are designed to save human lives, but we would not have any reason for such pro-attitudes if they had been poorly designed, so that they were likely to kill rather than to save people. (Nevertheless, we might still admire the intentions of the designers, even if we disapprove of their technical skills).

It is not difficult to find other examples that fit in this second category. Such examples include, for example, a safe chemical plant, a sustainable light bulb or a gender equitable computer game. If we call a chemical plant safe we do not merely mean that it is used in a safe way but rather that it is – if properly used – safe, for example in the sense of making accidents unlikely. We thus mean that the plant is designed for safety (although it will also be designed for other goals and values) and that it is actually likely to be safe in practice. Similarly a sustainable light bulb is not one that is used in a sustainable way, but rather one that – if used properly – does not consume a lot of energy and that has been intentionally designed for this feature. A gender equal computer game is a computer game that is intended to be interesting for and to meet the interests of boys and girls, men and women, and has designed features that make it possible to realize this. In these examples, safety, sustainability and gender equity are values that the artifact embodies on the basis of certain designed features, even if they do not refer to the function of the artifact. Safety is not the function of a safe chemical plant, nor is its function – producing certain chemical substances – conducive to safety. Similarly, it would be strange to say that the function of the computer game is gender equity. Nevertheless it may well be the case that the game is so designed that its designed features are conducive to gender equity. Examples like these show that it is possible to design for a (positive) extrinsic final value in other ways than incorporating this value in the artifact's function. This may also be achieved by designing a technical artifact for a value and by seeing to it that it has the appropriate designed features to realize this value.

## 7.8 The Importance of Design

The concluding observation of the previous section suggests the following general claim:[13]

[13] Our analysis shows that the following conditions are sufficient for embodying extrinsic final value; whether they are necessary conditions remains to be seen.

*The designed properties of a technical artifact x form the resultance base of an extrinsic final value G if the following two conditions are met:*

1. *The designed properties of x have the potential to achieve or contribute to G (under appropriate circumstances)*
2. *x has been designed for G*

We discussed both conditions for the class of examples in which the embodied value of an artifact does not coincide with its function. It is easy to see that the conditions also apply if G is part of the function of a technical artifact. On the dual nature account of technical artifacts, for F to be the function of a technical artifact x, it is minimally required that (1) F was intended by the designers to be the function of x, i.e. that the designers purposively designed x for F and (2) x has the capacity to realize F in the appropriate circumstances. These conditions entail the above mentioned conditions if G is part of, or identical to, F. Somewhat analogous to the dual nature account, the embodiment of extrinsic final values in technical artifacts thus depends on both an intentional condition ('x has been designed for G') and on a condition that primarily refers to physical properties ('The designed properties of x have the potential to achieve or contribute to G (under the appropriate conditions)').[14]

The phrase 'x has been designed for G' can mean a number of things here. Minimally it means that efforts have been made to design x so that it has the capacity to be conducive to G in the appropriate circumstances. In addition, it can also mean that x is optimized for G, or that efforts have been made in the design process to prevent uses of x that would destroy (or otherwise express a negative attitude towards) G, or it can mean that efforts have been made to make x fit for the circumstances in which it is (usually) appropriate to express a pro-attitude towards G. It should be noted that 'x has been designed for G' does not necessarily mean that x has been designed according to the approach of Value Sensitive Design (VSD) as this approach has recently been advocated by a number of authors. In our opinion, design for values is much older than the recent attention for VSD suggests. It is what many designers have been doing all the time. Design for values is thus probably as old as designing itself (although the emphasis on designing for moral values may be a recent phenomenon).

Back to our central issue: Is the above result a rebuttal of N3? Only in so far as it can be shown that indeed artifacts can be designed such that they fulfill the above two conditions. In the previous section we have discussed a variety of examples satisfying both conditions and it is not difficult to provide many other ones. In the light of our original question, whether VSD is possible, the second condition ('x has been designed for G') may seem a bit paradoxical or even question-begging because it sounds like VSD is possible just by trying. This is, however, not true because the first condition requires that not just an attempt is made but that the designed

[14] "Primarily" because the formulation leaves open that some of the designed properties are textual or symbolic. We take it, however, as characteristic for technical artifacts that their designed properties are by and large physical properties and that their symbolic/textual features are somehow related to the physical properties that are conducive to realizing their technical function.

properties have the potential to achieve or contribute to G (under appropriate circumstances). In the next section, we will say a bit more how this potential may be assessed in practical cases and how the phrase 'appropriate circumstances' may be understood. For the moment, we note that one might not just want to require that x is conducive to G under appropriate circumstances but that it is so because x has been designed for G, i.e. that 'x has been designed for G' is part of the explanation why 'the designed properties of x are conducive to a final value G.' Our final proposal therefore reads[15]:

*The designed properties of a technical artifact x form the resultance base of an extrinsic final value G, so that x embodies G, if the designed properties of x have the potential to achieve or contribute to G (under appropriate circumstances) due to the fact that x has been designed for G.*

## 7.9 Realized Versus Embodied Value

We have argued that it is possible to embody a specific kind of value, namely extrinsic final value, in technical artifacts through design. We want to stress, however, that an embodied value is not necessarily realized in practice. To see why, we have to recall the distinction between resultance base and supervenience base. Figure 7.1 clarifies the relation between what may be called *intended value* (the value which designers aim to embody in their design and which they hope to be realized in practice), *embodied value* and *realized value*. As this diagram suggests, use and the context of use are crucial for whether embodied value is indeed realized in practice.

Figure 7.1 raises the question whether we can somehow verify what value is embodied in a designed artifact. Is there any way of telling what value G, if any, is embodied in a designed technical artifact x? We can, at least to some extent, observe and experience values that are realized in user practices, but can we observe or experience embodied value? *Prima facie* the answer appears to be negative because embodied value is more like a capacity (a potential value), the actual realization of which depends on a broader supervenience base, including the context of use.

It may, however, be possible to infer the embodied value of a technical artifact from its realized value in various use contexts. In addition to such inferences, it might be possible to make embodied value more directly traceable by specifying it in a specific way, namely as that value that is realized if an artifact is properly used. The underlying idea is that designers often not just design an artifact but in doing so also design, or at least presuppose, a proper way to use the artifact. Proper use may,

[15] Our final proposal comes close to a suggestion made (but not further elaborated) by Franssen (2009: 947–948): "technical artefacts can be called bad in a moral sense if its functional requirements, the characteristics that in a sense define it, explicitly refer to specifically morally bad states of affairs as goals states to be realized by using the artifact, such that it will be optimized, through the accepted methods of engineering design, to realize precisely these outcomes."

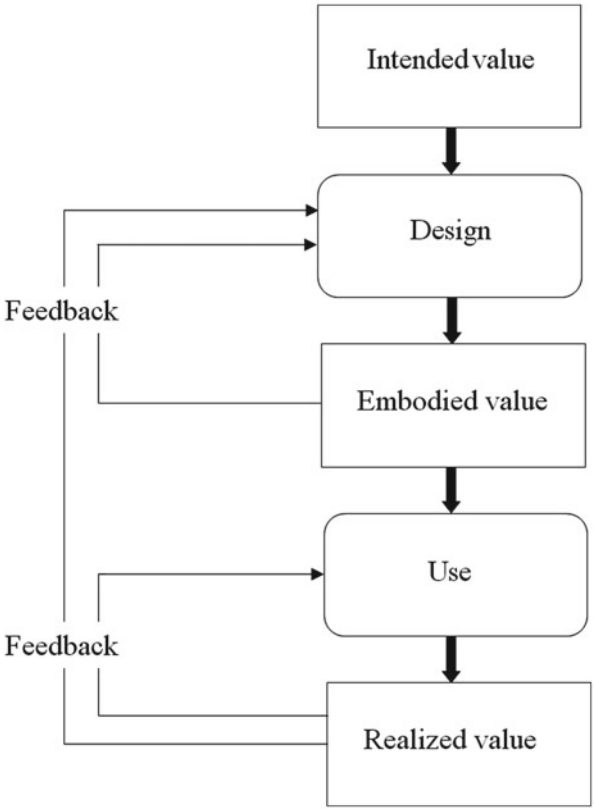


**Fig. 7.1** The relation between intended, embodied and realized value

for example, be defined by what Houkes and Vermaas call the use plan of a technical artifact (Houkes and Vermaas 2004, 2010). According to them, the design of technical artifacts is always also the design of a use plan.

The advantage of defining embodied value as the value that is realized if an artifact is properly used is that embodied value becomes more directly traceable and that engineers are better able to verify whether their designs embody the intended values. Two remarks are, however, to be kept in mind. First, proper use may underdetermine what value is realized. It is very well conceivable that proper use in different use contexts leads to the realization of (somewhat) different values. In other words, the supervenience base that determines what value is realized may be broader than the designed features and the features defined by proper use together. So the notion of proper use is not an error-free method for ascertaining the embodied value of a technical artifact, although it may be helpful. Second, the ultimate aim of approaches like VSD is to contribute to the realization of values in actual practices. From the view point of VSD, embodying values in artifacts, in the sense we have defined the term here, is only a first step. It is, for example, conceivable that designers are successful

in embodying a value in a technical artifact by articulating a rather eccentric or unrealistic form of proper use for that artifact. In such cases their 'success' seems hardly relevant for the ultimate goal of VSD, i.e. realizing values in practice.

This brings us to a final point. In our opinion it is part of the responsibility of designers to try to anticipate the circumstances and ways in which artifacts will be used and to try to anticipate how this will affect the realization of values. This is not to say that designers should always accept current user practices. They may sometimes have good reasons to ask users to 'properly use' an artifact in a way that deviates from what they are used to. In other cases, however, it might be that the designers have to adapt their notion of 'proper use' to actual practices or to what can be realistically expected from users. We also do not want to suggest that designers can precisely predict or control how artifacts will be used and what values will be realized in practice (cf. Albrechtslund 2007). We nevertheless think that fruitful design for values requires that designers try to anticipate actual use and the actual realization of values. Moreover, they ought to monitor whether values are realized in practice and feed such insights back into the design process.

In summary, the central outcome of our analysis is that the neutrality thesis does not hold and that it is possible for technical artifacts to embody values. However, the values that may be embodied in technical artifacts are of a specific kind, namely extrinsic final values. Values may be designed into technical artifacts and therefore VSD is possible. We have also briefly argued that the main difficulty that VSD faces is not embodying values in technical artifacts through design, but that the real challenge for VSD lies in realizing such embodied values in actual use practices.

**Acknowledgement** Ibo van de Poel is grateful to NIAS, the Netherlands Institute for Advanced Study, for providing him with the opportunity, as a Fellow-in-Residence, to write this paper.

## Appendix: The Instrumental Value of Technical Artifacts

Phrases like 'x is a good knife' refer to goodness of x as an instance of a kind, in this case goodness as a knife. Usually this goodness is understood as a kind of instrumental goodness. The underlying idea is that kinds of technical artifacts can be associated with certain purposes or certain functions for which they have been designed. So if we say that 'x is a good knife' that can be analyzed as saying that 'x is a knife' and that, assuming that the function of knives is cutting, 'x is good for cutting.' The latter statement refers to instrumental value. Now if we want to know whether this instrumental value is really a value at all, we can employ thesis V (see main text): if instrumental value is real value it should correspond with reasons for a positive attitude towards the instrumentally valuable object and these reasons should originate from the same resultance base as the instrumental value itself. But does it? In answering this question we start with the account Maarten Franssen has developed to characterize the normativity of

evaluative statements such as 'this is a good knife', i.e., evaluative statements about the goodness of technical artifacts as instruments. He proposes the following characterization of such evaluative statements:

(1) 'x is a good K' expresses the normative fact that x has certain features f that make x a K and that make it the case that a person p's wish to K recommends that p uses x for K-ing. (Franssen 2009: 933)

Here K refers to a certain type of technical artifact (like a knife), and x refers to a token of this artifact type; K-ing is the use or performance of the function of a K (cutting in the case of knifes). f is what we have called in the main text the resultance base for the instrumental value and the reasons or recommendations are associated with this value. The term recommendation refers to what Broome (1999) has called a normative recommendation: 'x recommends y for p' means that 'p has reason to see to it that (if x is the case then y is the case)'.

Franssen also addresses the question whether the instrumental value of x is really a value. His suggestion is that while the instrumental value of x may give us reasons to use x, using is really not the expression of a pro-attitude. Since to have value corresponds with reasons for a positive response (a pro-attitude or a pro-behavior) as expressed in V, it follows that x does not have value because it gives reason for using, since using is not a positive response according to Franssen.

The argument that using is not a positive response, however, appears not very convincing. After all using an artifact costs efforts and doing so therefore may be taken to imply somehow a positive response. Moreover 'x is a good K' seems not only to recommend that 'p uses x for K-ing if p wishes to K', but also that 'p keeps (instead of throwing away), maintains or even buys x for K-ing if p wishes to K' because all these activities enable or ensure that p can use p for K-ing. Keeping, maintaining and buying seem all pro-behaviors expressing a positive attitude. However, even if using is considered to be a pro-attitude, there may be another way to understand why the instrumental value of an artifact is not a value at all, namely that it does not correspond with reasons, or at least not with reasons of the right kind (i.e. reasons originating from the artifact itself).

The normative recommendation that is expressed in (1) is equivalent to a reason 'to see to it that (if p wishes to K, then p uses x for K-ing)'. This reason, however, is not grounded in the (instrumental) value of x, but rather in the rationality requirement or recommendation that if one wishes something one should (or is recommended to) adopt appropriate means to achieve it. In as far as (1) expresses certain reasons these reasons are grounded in (the value of) rationality, rather than in the specific value of x. Another way of seeing this is to recognize that if p has no reason for wishing to K, p also has no reason to use x for K-ing. The mere fact that 'p wishes to K' cannot give p any reason to K (at least according to such authors as Raz 1986; Scanlon 1998; Dancy 2002). So, in as far as (1) gives reasons to use x it are the wrong kind of reasons for V because it are reasons not grounded in the valuable object (they have another resultance base than f) and hence the value of the object cannot be associated with those reasons. Therefore the instrumental value that is expressed in (1) does not give a reason for a positive response to x.

## References


Albrechtslund, A. (2007). Ethics and technology design. *Ethics and Information Technology, 9*, 63–72.

Anderson, E. (1993). *Value in ethics and economics*. Cambridge, MA: Harvard University Press.

Broome, J. (1999). Normative requirements. *Ratio, 12*, 398–419.

Dancy, J. (2002). *Practical reality*. Oxford: Oxford University Press.

Dancy, J. (2005a). The particularist's progress. In T. Rønnow-Rasmussen & M. J. Zimmerman (Eds.), *Recent work on intrinsic value* (pp. 325–347). Dordrecht: Springer.

Dancy, J. (2005b). Should we pass the buck? In T. Rønnow-Rasmussen & M. J. Zimmerman (Eds.), *Recent work on intrinsic value* (pp. 33–44). Dordrecht: Springer.

Feldman, F. (2005). Hyperventilating about intrinsic value. In T. Rønnow-Rasmussen & M. J. Zimmerman (Eds.), *Recent work on intrinsic value* (pp. 45–58). Dordrecht: Springer.

Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying values in technology. Theory and practice. In J. Van den Hoven & J. Weckert (Eds.), *Information technology and moral philosophy* (pp. 322–353). Cambridge: Cambridge University Press.

Franssen, M. (2009). Artefacts and normativity. In A. Meijers (Ed.), *Handbook of the philosophy of science: Vol. 9: Philosophy of technology and engineering sciences* (pp. 923–952). Oxford: Elsevier.

Friedman, B. (1996). Value-sensitive design. *Interactions, 3*, 17–23.

Friedman, B., & Kahn, P. H., Jr. (2003). Human values, ethics and design. In J. Jacko & A. Sears (Eds.), *Handbook of human-computer interaction* (pp. 1177–1201). Mahwah: Lawrence Erlbaum Associates.

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

Herbich, J. B. (Ed.). (1999). *Handbook of coastal engineering* (Vol. N). New York: McGraw-Hill.

Houkes, W., & Vermaas, P. E. (2004). Actions versus functions. A plea for an alternative metaphysics of artefacts. *The Monist, 87*, 52–71.

Houkes, W., & Vermaas, P. E. (2010). In P. E. Vermaas (Ed.), *Technical functions: On the use and design of artefacts* (Vol. 1). Dordrecht: Springer.

Joerges, B. (1999). Do politics have artefacts? *Social Studies of Science, 29*, 411–431.

Kagan, S. (2005). Rethinking intrinsic value. In T. Rønnow-Rasmussen & M. J. Zimmerman (Eds.), *Recent work on intrinsic value* (pp. 97–114). Dordrecht: Springer.

Korsgaard, C. M. (1983). Two distinctions in goodness. *Philosophical Review, 92*, 169–195.

Kroes, P. (2010). Engineering and the dual nature of technical artefacts. *Cambridge Journal of Economics, 34*, 51–62.

Kroes, P. (2012). *Technical artefacts: Creations of mind and matter: A philosophy of engineering design*. Dordrecht: Springer.

Kroes, P., & Meijers, A. (2006). The dual nature of technical artefacts. *Studies in History and Philosophy of Science, 37*, 1–4.

Mitcham, C. (1994). *Thinking through technology. The path between engineering and philosophy*. Chicago/London: University of Chicago Press.

Moore, G. E. (1903). *Principia ethica*. Cambridge: Cambridge University Press.

Moore, G. E. (1912). *Ethics*. Oxford: Oxford University Press.

Moore, G. E. (1922). The conception of intrinsic value. In *Philosophical studies*. New York: Harcourt, Brace.

Pitt, J. C. (2000). *Thinking about technology. Foundations of the philosophy of technology*. New York: Seven Bridges Press.

Rabinowicz, W., & Rønnow-Rasmussen, T. (2005). A distinction in value: Intrinsic and for its own sake. In T. Rønnow-Rasmussen & M. J. Zimmerman (Eds.), *Recent work on intrinsic value* (pp. 115–129). Dordrecht: Springer.

Raz, J. (1986). *The morality of freedom*. Oxford: Oxford University Press.

Raz, J. (1999). *Engaging reason. On the theory of value and action*. Oxford: Oxford University Press.
Rønnow-Rasmussen, T. (2002). Instrumental values – Strong and weak. *Ethical Theory and Moral Practice, 5*, 23–43.
Ross, W. D. (1930). *The right and the good*. Oxford: Clarendon Press.
Scanlon, T. M. (1998). *What we owe to each other*. Cambridge, MA: Harvard University Press.
Schroeder, M. (2009). Value theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2008 ed.). http://plato.stanford.edu/archives/fall2008/entries/value-theory/
Snippen, E., Barneveld, H. J., Flikweert, J. J., & Timmer, D. F. (2005). The role of guidelines in safety against flooding. In J. Van Alphen, E. van Beek, & M. Taal (Eds.), *Floods. From defense to management* (pp. 701–705). London: Taylor & Francis.
Van Eck, D. (2011). *Functional decomposition: On rationality and incommensurability in engineering, TPM: Section philosophy*. Delft: Delft University of Technology.
von Wright, G. H. (1963). *The varieties of goodness*. London: Routledge & Kegan Paul.
Winner, L. (1980). Do artifacts have politics? *Daedalus, 109*, 121–136.

# Chapter 8
# From Moral Agents to Moral Factors: The Structural Ethics Approach

**Philip Brey**

**Abstract** It has become a popular position in the philosophy of technology to claim that some or all technological artifacts can qualify as moral agents. This position has been developed to account for the moral role of technological artifacts in society and to help clarify the moral responsibility of engineers in design. In this paper, I will evaluate various positions in favor of the view that technological artifacts are or can be moral agents. I will find that these positions, while expressing important insights about the moral role of technological artifacts, are ultimately lacking because they obscure important differences between human moral agents and technological artifacts. I then develop an alternative view, which does not ascribe moral agency to artifacts, but does attribute to them important moral roles. I call this approach structural ethics. Structural ethics is complementary to individual ethics, which is the ethical study of individual human agents and their behaviors. Structural ethics focuses on ethical aspects of social and material networks and arrangements, and their components, which include humans, animals, artifacts, natural objects, and complex structures composed of such entities, like organizations. In structural ethics, components of networks that have moral implications are called moral factors. Artifact ethics is the study of individual artifacts within structural ethics. It studies how technological artifacts may have a role as moral factors in various kinds of social and material arrangements as well as across arrangements. I argue that structural ethics and artifact ethics provide a sound alternative to approaches that attribute moral agency to artifacts. I end by arguing that some advanced future technological systems, such as robots, may have capacities for moral deliberation which may make them resemble human moral agents, but that even such systems will likely lack important features of human agents which they need to qualify as full-blown human agents.

P. Brey (✉)
Department of Philosophy, University of Twente, Enschede, The Netherlands
e-mail: P.A.E.Brey@utwente.nl

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_8,

## 8.1 Introduction

Recently, a number of authors in the philosophy of technology have argued that some or all technological artifacts can qualify as moral agents. The notion of a moral agent has traditionally been reserved for human beings, and is used to refer to beings which can be held morally responsible for their actions. Such beings have the capacity to know right from wrong and are able to choose their actions freely based upon their considered moral judgments. Yet, some authors have argued, extending the notion of moral agency to technological artifacts is necessary in order to account for the moral role of (some) artifacts, which is in some cases highly similar to that of human agents. In addition, they have argued, doing so will be useful for the attribution of moral responsibility to designers.

In this paper I will evaluate various positions in favor of the view that technological artifacts can be moral agents. I will find that these positions bear important insights about the moral role of technological artifacts, but are ultimately lacking. I then develop an alternative view, which does not ascribe moral agency to artifacts, but does attribute to them important moral roles. I call this approach structural ethics. I will argue that this approach has all the benefits of approaches that ascribe moral agency to artifacts, while maintaining a distinction between the moral agency of humans and the moral roles of nonhuman entities like technological artifacts.

## 8.2 The Philosophical Concept of Moral Agency

To begin my inquiry, I will give an account of the classical notion of a moral agent as it has been developed in philosophy. In the next section, this account will then be contrasted with extended notions of moral agent that have been developed in the philosophy of technology. The standard notion of a moral agent is a philosophical notion that refers to beings which are capable of acting morally and are expected by others to do so. Although there is no generally agreed definition of "moral agent," existing definitions tend to emphasize three features.[1] Moral agents are beings that are (1) capable of reasoning, judging and acting with reference to right and wrong; (2) expected to adhere to standards of morality for their actions; and (3) morally responsible for their actions and accountable for their consequences. These three features together define what I will call the standard philosophical conception of a moral agent, or, in brief, the *standard conception*.

Which beings qualify as moral agents on the standard conception? Given the three mentioned features, it appears clear that only adult, rational human beings do. Only rational human beings are capable of moral reasoning, and only they are expected to behave morally and are held morally accountable. Adults that are incapable of distinguishing right from wrong are not normally seen as moral

[1] See Himma (2009) for some definitions of "moral agent".

agents, and are not held to be morally responsible for their actions. Similarly, young children are not held to be moral agents, nor are animals or inanimate objects. In contrast, we expect "normal" adults to have a developed capacity for moral reasoning, judgment and action, we expect them to exercise that capacity, and we hold them accountable if they nevertheless engage in immoral acts. In short, we hold them to be moral agents.

A moral agent is a special kind of *agent*. An agent, in the philosophical sense, is a being capable of performing *actions*. Actions constitute a special class of behaviors, since not any kind of behavior constitutes an action (Davidson 1980). Breathing, for instance, is not an action, even if it is something we do. Actions are intentional, they depend on capacities for rational thought and self-interested judgments, and the performance of goal-directed behaviors based on such thoughts and judgments. Typically, actions are explained by reference to *intentional states* of the agent. Intentional states are mental states like beliefs, desires, fears, perceptions and intentions, that have a directedness to something else. For example, an explanation of why John drank the milk (an action) would refer to intentional states of John that explain his behavior, e.g., John's fear that he was dehydrating and his belief that the milk would quench his thirst. In contrast, a mere behavior (e.g., John's blinking, or his falling when pushed over) would refer to mere physical causes like sand getting into John's eye or someone shoving John. Agents, in conclusion, are beings capable of performing actions, which are behaviors caused by intentional states of goal-directed beings.

An agent is a *moral* agent when the intentional states that it cultivates and the subsequent actions it performs are guided by moral considerations. This requires, first of all, a capacity for *moral deliberation*, which is reasoning in order to determine what the right thing to do is in a given situation. A capacity for moral deliberation requires a capacity for reasoning and knowledge of right and wrong. Moral deliberation typically results in *moral judgments*, which are judgments about right and wrong. It also frequently results in intentions to perform certain actions that are held to be morally good, and to refrain from performing actions that are held to be immoral. For example, a moral agent may deliberate on what to do with a found wallet, in a way that takes into account both moral and non-moral considerations. He may then arrive at the moral judgment that turning the wallet in to the police is the right thing to do. This may then result in an intention to give the wallet to the police, and a subsequent action of giving the wallet to the police.

Let us now turn to the second feature of moral agents, which is that they are beings that are expected to behave morally. This is a *normative* rather than a *factual* expectation. That is, we believe that people *should* behave morally, that they have a *moral obligation* to do so. We do not expect that they in fact always do. In fact, we know that they often do not. However, our belief in morality, and our knowledge that others are capable of moral actions, results in an expectation that others behave morally, or at least make every effort to do so. We do not find it acceptable that people either do not engage in moral deliberation in situations that pose moral dilemmas, or do so but nevertheless act immorally.

The third feature of moral agents, their being held morally responsible and accountable, is a corollary of the first and second feature. Because people are capable of acting morally, and because we expect them to do so, we hold them to be *morally responsible* for their actions. That is, we hold that their actions are appropriately the subject of moral evaluation by others, and of particular kinds of reactions based on such moral evaluations. Such reactions particularly include praise and blame, as well as related responses such as reward and punishment. Thus, if someone acts morally, we have a propensity to praise them for doing so, whereas if someone acts immorally, we may blame or condemn them for their actions. Moral responsibility is usually held to presuppose *free will*: persons have moral responsibility to the extent that they can freely choose their acts. *Moral accountability* is a type of moral responsibility that goes beyond it in assuming the existence of shared moral standards in a community that can be alluded to in evaluating someone's actions (Watson 1996). When there are such shared standards, moral agents can be praised or blamed with explicit reference to such interpersonal standards. They can be judged to either have upheld or have broken these standards, and can be held accountable for doing so.

## 8.3 Theories of Artifacts as Moral Agents

Given the prevailing conception of a moral agent in philosophy, it would seem unlikely that anything else but a human being could qualify as a moral agent. It seems particularly unlikely that technological artifacts like lawnmowers and iPods could qualify as moral agents. Technological artifacts are not capable of moral deliberation, they are not expected to behave morally, and they are not held to be morally responsible or accountable. They therefore seem very poor candidates for moral agents. Recently, however, several philosophers have defended the view that the notion of a moral agent should be extended to include technological artifacts.

There are two versions of this view, which I will now lay out. On the first view, which I will call the *moral artifacts view*, all technological artifacts are, or could function as, moral agents. This view was first proposed, although not in very explicit terms, by Bruno Latour (1992). It has subsequently been defended by Powers and Johnson (2004), Keulartz et al. (2004) and Verbeek (2008). On the second view, which I will call the *morally intelligent agents* view, certain highly evolved technological artifacts, namely those capable of autonomous behavior and intelligent information processing, qualify as moral agents. On this view, the class of moral agents includes, next to human beings, things like autonomous robots and software agents. This view was first proposed by Allen et al. (2000), and has subsequently been developed in an influential article by Luciano Floridi and Jeff Sanders (2004). It has also been defended by a number of other authors, including Stahl (2004), Sullins (2006) and Johnson and Powers (2008).

The moral artifacts view asks for a major revision of our concept of moral agency, extending it to mundane artifacts like knives, automobiles and bridges. The second

approach asks for a more limited revision of the standard conception. In this paper, my focus will be on the more radical claim, which is the moral artifacts view. Near the end of the paper, I will also briefly go into the moral intelligent agents view.

The moral artifacts view rests on the observation that the roles that technological artifacts play in human affairs are frequently not morally neutral. There seem to be two ways in which technological artifacts can have a moral impact. First, artifacts are capable of *steering moral behavior in humans*. Artifacts may stimulate or enforce morality by making humans behave morally. For example, a car that flashes a warning when driver or passengers do not wear a seat belt stimulates moral behavior by its users. Second, artifacts are capable of *influencing moral outcomes*. Even when artifacts do not influence moral behavior in humans, they may have consequences that can be morally evaluated. For example, a computer network that randomly provides added bandwidth to some of its users is less just than a computer network that gives all its users equal bandwidth. The network that provides equal bandwidth hence generates a better moral result, even though it does not influence any person to behave morally. Both of these moral roles of artifacts have been used to argue that artifacts are, or can be, a type of moral agent.

Bruno Latour (1992) arrives at the view that artifacts are moral agents by asking the question whether morality is only located in people or also in things. He argues that moral laws in a society are not only enforced by humans but also by artifacts. Artifacts make up the "missing masses" that together with humans make up the moral fabric of society. Artifacts enforce moral rules in a way that is similar to that of humans, Latour argues. For instance, a moral (and legal) rule that tells us to drive slowly in a densely populated neighborhood can be enforced either by a police officer who waves cars down, a street sign that tells drivers to slow down, or a speed bump that forces them to slow down. Morality is hence similarly enforced by both humans and artifacts.

Latour argues that both humans and artifacts are bearers of *programs of action* that aim to enforce particular moral or social rules or configurations. A police officer and a speed bump may, for example, both aim to enforce the rule "IF a car drives in this street, THEN its speed is no more than 30 m.p.h." Bearers of such programs of action are called "agents" or "actants" by Latour. Artifacts, on his view, are *moral* agents when they are bearers of a program of action that enforces a moral rule. Typically, such programs of action are inscribed into the design of an artifact. Latour claims that many mundane artifacts enforce moral rules by facilitating, stimulating, or forcing behaviors and situations that comport with everyday morality.

Powers and Johnson (2004) present an alternative account that revolves around the notion of intentionality. They define a view of agency according to which causality and intentionality, but not mentality, are necessary features of it. Artifacts, they observe, are causally efficacious, meaning that their presence and operation has consequences for what happens in the world. In addition, they argue, artifacts are bearers of intentional states. This claim is based on the observation that artifacts have a directedness at phenomena external to them. For instance, a telephone is directed at human fingers and ears and at human verbal communication. These different types of directedness of telephones constitute different intentional states,

according to Powers and Johnson. The intentionality of artifacts is bound up with their function, which defines their intended use, and which is a result of the intentions of designers.

According to Powers and Johnson, the intentional states of artifacts allow for reason explanations instead of mere causal explanations of events. For example, if a speed bump slows down a car, its slowing down can be explained by reference to the directedness of speed bumps at cars and their function of slowing cars down, both of which were intended by a designer. A speed bump is hence different from a pothole, which merely happens to cause a car to slow down. On their view, the speed bump is therefore an agent, whereas the pothole is not. A speed bump is moreover a moral agent because it enforces moral rules and has consequences for moral patients. Because artifacts have intentional states and have moral consequences, it is also possible to make attributions of moral responsibility, Powers and Johnson argue. Moral responsibility for the agency of an artifact lies with the human agents who put its intentional states into it, including, most prominently, the designers.

While both Latour and Powers and Johnson conceive artifacts as agents, they also emphasize that artifacts cannot perform any actions independently of users and designers, that is, of human agents. Powers and Johnson emphasize that "the behavior that results [from artifacts] is a combination of the intentionality of the artifact designer, input from the user, and behavior of the artifact" (2004: pp. 22–23). And Latour emphasizes, similarly, that artifacts do not generate moral outcomes by themselves. For instance, it is not just the flashing light in the car that causes one to wear a seat belt: "I, plus the car, plus the dozens of patented engineers, plus the police are making me be moral" (1992: p. 226).

If artifacts are always dependent on human agents like users and designers for their agency, does this not demonstrate an asymmetry between human and nonhuman moral agency? Isn't it the case that humans are able to perform moral acts autonomously, whereas artifacts are always dependent on human agents? It would seem that Powers and Johnson are willing to accept this asymmetry. Latour and his followers, however, do not. On Latour's view, human and nonhuman agents are both dependent on constellations or networks of agents for their actions. These networks, which Latour calls actor-networks, consist of both human and nonhuman agents (Latour 1987). Human agency is, on Latour's view, always the product of multiple agents co-acting with a human agent. Just like the car does not act alone in causing me to wear a seat belt, I do not act alone in wearing the seat belt. My action is caused not just by me but also by the blinking light in my car, the designers behind it, and the police that checks on seat belt use.

Latour hence does not only extend the notions of agency and moral agency to artifacts, he also engages in a major revision of the concept of human agency. Human agency is, on his view, not attributable to agents, but is rather a property of actor networks, in which multiple actors together produce a particular action. Attributing an action to a particular actor (human or nonhuman) is merely a matter of putting the focus on that actor, while we might have also chosen to emphasize the role of other actors in the network. Morality, in this view, is similarly a property of networks consisting of human and nonhuman entities that together generate moral

actions and moral outcomes. This position has been further defended by Keulartz et al. (2004) and by Verbeek (2005, 2008), who however rejects the ontological symmetry between people and things proposed by Latour.

We hence have seen several arguments for extending the notion of moral agent to include technological artifacts. The authors who extend the notion in this way have several motives for doing so. They want to give greater visibility to the moral role of artifacts, to better account for the way morality is realized in society, and to allow for better, more ethical design and use of technological artifacts. In the next section, I will evaluate these arguments and discuss whether they provide sufficient reason to broaden the notion of moral agent to include technological artifacts.

## 8.4 Evaluating the Moral Artifacts View

Proponents of the moral artifacts view present novel conceptions of moral agency that are intended to replace rather than supplement the existing philosophical concept of moral agency. Most authors do not hold that their view is necessarily ontologically more correct, but rather emphasize its pragmatic usefulness in understanding the moral role of artifacts and their relation to humans. Thus, Powers and Johnston say that they have "practical reasons for calling technological artifacts agents" and use this terminology to "highlight that the ways in which artifacts are designed and used have powerful consequences for the moral character of the world we inhabit" (2004: p. 26). Similarly, Keulartz et al. say that they think that "it is useful to speak of artifacts as (possible) moral agents. Not for ontological reasons, but for pragmatist ones" and say that this conceptualization highlights important aspects of the relations between humans, technological artifacts and ethics.

I agree with these authors that concepts should primarily be evaluated on pragmatic rather than ontological grounds. As Wittgenstein, Peirce and others have shown, we do not usually use concepts to describe objective essences, but rather to selectively highlight aspects of things that are important to us in dealing with them. Consequently, a concept is a good (i.e., useful) concept if it highlights important aspects of a thing or state-of-affairs while not obscuring important other ones. So the question for the moral artifacts view is whether it (a) highlights important phenomena that were previously overlooked, and (b) does not obscure other important phenomena.

The main benefit of the moral artifacts view is that it highlights the facts that technological artifacts play an important role in shaping moral actions and outcomes and that they are part of the moral fabric of society. Technological artifacts have been largely overlooked in moral theory, and have only been assigned an instrumental role in human action, as means that make certain actions possible, or make them easier to perform. Because of this instrumental conception, artifacts are normally thought of as morally neutral. All morality is thought to be located in the choosing and acting human subject. Yet, as Hans Jonas has argued, technological artifacts drastically change human action, and this has important consequences for

ethics (Jonas 1984). The moral artifacts view helps us arrive at a better view of technological artifacts that reveals their important role in shaping moral action and moral outcomes.

Another benefit of the moral artifacts view is that it highlights useful similarities between human agents and artifacts regarding their moral role. As Latour has shown, both human agents and artifacts can be used to enforce the same moral norms, both can influence humans to behave morally, and both can determine moral outcomes. Both, in addition, are dependent on other entities, both human and non-human, for being able to play these roles. These important similarities between them have been less obvious in traditional accounts. In addition, as Powers and Johnson have shown, both human agents and artifacts exhibit intentionality. The intentionality in artifacts can be referred to in making reason explanations, or intentional explanations, of moral outcomes, just as it can in humans. It can even be used to make attributions of moral responsibility, just as it can in humans, by linking the consequences of artifacts to the designers who inscribed their intentionality into these artifacts.

These are great benefits of the moral artifacts view. However, I will argue, this view also obscures and obliterates important phenomena. First, it obscures differences between human agency and the agency of artifacts and the unique characteristics of human agency and human action. Actions, unlike mere events, result from the intentional states of goal-directed, interest-bound beings who intend to cause changes in the world, and these intentional states provide reasons for the action in question. Artifacts are not normally goal-directed, they do not have interests, and they do not have intentional states like beliefs and desires that cause their behavior. Replacing the standard conception of agency with an extended one means that these important differences are obscured.

There are at least two reasons why obscuring these differences is a bad idea. First, the classical notion of an agent has an important role in our moral image of a human being. Part of what makes human beings special and valuable is their ability to form intentional states like beliefs and desires, and then choose to act according to them. In this, they differ from things like rocks and coconuts, which can only passively cause things to happen, without reason or intent. As soon as things like screwdrivers are also called agents, these special features of human agency are lost in our understanding of agency, and the moral image of humans is damaged as a result.

Second, the classical notion of agency has an important role in explaining and accounting for events. Actions, and any events following from them, are explained by reference to reasons and intentions, unlike most other events, which are explained by reference to mere causes. Reason explanations provide us with different information than causal explanations. They give us insights into the motives and justifications of human agents. Our responses to actions tend to be different from those to mere causal events: we tend to the beliefs, desires, and other mental states that underlie these actions, and we do not only respond physically, but also morally and socially. However, an extended notion of agency obliterates this distinction between actions and mere events, and hence the special role of actions in our understanding of the world.

Against this point, Powers and Johnson may object that on their account of agency, all actions rest on intentional states and can be explained intentionally. So their account at least seems to preserve the difference between actions and mere events, and hence between intentional and causal explanation. I believe their account is flawed, however, by attributing intentional states to artifacts. Artifacts certainly have intentional *properties*. For example, a speed bump has a directedness to automobiles that it has been designed to slow down. This is an instance of what John Searle (1984) has called *derived intentionality*: a directedness deriving from human intentions that have been inscribed into artifacts. But artifacts do not have intentional *states*. That is, they do not have, as humans do, states like beliefs, desires, and intentions that provide reasons for their actions. It is false to say: "The car slowed down because the speed bump intended it to slow down".

However, one can correctly say either "The car slowed down because it is the *function* of speed bumps to slow down cars" or "The car slowed down because speed bumps *are intended* to slow down cars". The former explanation is a functional explanation rather than an intentional explanation, and does not require any attribution of intentional states to artifacts. The latter is an intentional explanation, but it is left implicit who is doing the intending. Surely, however, it is not the speed bump which is doing the intending. Rather, it is the designers and traffic controllers who intend the cars to slow down *by means of* a speed bump. So a full intentional explanation would read: "The car slowed down because speed bumps are intended by designers and traffic controllers to slow down cars". But this account also does not require any attribution of intentional states to artifacts. Rather, it seems accurate to say that the intentional states belong to the designers and traffic controllers, and it is their actions (the development and installation of speed bumps) that cause cars to slow down.

Next to obliterating the distinction between agents and mere inanimate objects, the moral artifacts view also obscures the difference between moral agents and entities that have mere moral properties or implications. Most importantly, what is lost in the equivocation is the idea of a moral agent as an agent capable of moral deliberation and of actions based on such deliberation, and the idea of a moral agent as a morally responsible and accountable being.

The capacity for moral deliberation in human moral agents is important because it enables a very different mode of interaction than is possible with entities that lack this capacity. Things that lack this capacity but do play a moral role, like speed bumps, can only be interacted with physically. Speed bumps can be physically created, redesigned or removed. We can have a similar physical mode of interaction with human beings. However, because humans engage in moral deliberation, we can also enter into verbal modes with them: we can deliberate with them, bring forward arguments or ideas, try to convince them, threaten them or influence them otherwise. In attempts to influence moral behavior and moral outcomes, a physical mode of interaction is often the last one we choose with human beings. This is because they are sentient beings capable of moral deliberation. This important capacity is obscured, however, when it is no longer held to be a defining property of moral agents.

Removing the notion of moral responsibility from our conception of moral agency is also unappealing. The concept of moral responsibility is important to us because we believe that people should accept that their actions are subjected to moral standards, that they should be able to defend the moral rightness of their actions to others, and that others can appropriately respond to their actions with their own attitudes, judgments and actions that include praise, blame, punishment and reward. Philosophers have put forward two different reasons why such praise and blame (and punishment and reward) should be issued (Eshleman 2009). The first, expressed by the merit-based view of moral responsibility, is that praise and blame should be issued to moral agents because they deserve such responses from others. Those who act immorally deserve blame and punishment, and those who act morally deserve praise and reward. The second, encoded in the consequentialist view, is that praise and blame should be issued to moral agents in order to encourage future moral behavior and to prevent immoral behavior.

If we change from the standard view of moral agency to a broad view that includes moral artifacts, then notions like intentional action, moral deliberation and moral responsibility are no longer defining features of our notion of moral agency. This, I have tried to argue, is a significant loss. It may be argued that we could still retain these notions and attach them whenever the moral agents we refer to are human. This, however, is an insufficient response. Notions like those of agent and moral agent are fundamental concepts philosophers (and non-philosophers) use to understand and reason about reality. If these notions are restructured so as to lose important features, then these features are obscured in our understanding of reality. They are no longer activated whenever these concepts are activated in our minds, and as a result become less central in our thinking.

For the reasons given above, the gains brought by the moral artifacts view to include them are hence offset by considerable losses that result from important features of the standard conception of moral agency being obscured. As a result, this does not make the moral artifacts view particularly appealing. At the same time, the standard conception of moral agency also has its disadvantages, because it has tended to be accompanied by an instrumentalist understanding of technological artifacts that downplays their moral importance and does not reveal the similar roles that human agents and artifacts often play in giving shape to morality.

One way out may be to acknowledge the special role of humans, including their abilities of deliberation, while still maintaining that agency can be attributed to artifacts. This is the position taken by Verbeek (2005, 2008). Verbeek follows Latour in holding that agency is foremost a property of networks, but argues that humans have a central role in realizing agency and artifacts a contributory role. Artifacts mediate and hence co-construct human deliberations, intentions and actions. Human agency rarely, if ever, exists in pure form but is mediated by artifacts with which humans form human-artifact assemblies. Artifacts therefore cannot function as agents independently of humans, but they have agency in the sense that they contribute to the production of agency in human-artifact assemblies.

While I agree with Verbeek that human agency is often influenced by artifacts, and I am even willing to agree that agency can be attributed to human-artifact

assemblies, it does not follow that artifacts therefore have some form of agency, as Verbeek sometimes claims. This is like saying that because salty water is liquid, and it includes salt, that therefore salt has properties of a liquid. It would seem more correct to say that the salt in salty water mediates or transforms the liquidity of water without having liquid properties itself. To the extent that Verbeek assigns an independent intentionality to artifacts, his position will run into the same problems as the Powers-Johnson view. In addition, Verbeek has not demonstrated that human agency cannot exist independently from artifacts, and that therefore human beings cannot be conceived of as (moral) agents independently of the artifacts they use. Surely, it would seem, human beings that are bereft of any artifacts can still deliberate, intend or act. Humans are moral agents that continually couple with and decouple from artifacts that co-constitute their agency. Verbeek's view therefore gives too much credit to artifacts in assigning agency to them and too little to humans in denying them agency independent of, and prior to, any artifacts they may use.

I therefore conclude that another view is needed, one that incorporates the benefits of the standard conception of moral agency as well as those of the moral artifacts view, and does so without having significant drawbacks. It is to such a view that I will now turn.

## 8.5 An Alternative Account

What we have seen is that on the one hand, there are good reasons to retain the traditional notion of moral agent in its basic form and that on the other hand, there are also good reasons to upgrade the role of both technological artifacts and networks in ethics. My proposal is to introduce a new type of ethics, *structural ethics*, next to the familiar *individual ethics* that focuses on human (moral) agents. Structural ethics focuses on the moral aspects of social and material arrangements (structures or networks consisting of humans and nonhumans), including their impact on the actions of human agents. Structural ethics is intended to be complementary to individual ethics. Individual ethics is solely focused on the morality of individual human agents, their actions, and the intentional states and deliberations underlying them. As I will argue, structural ethics requires a new ethical vocabulary that is different from that of the moral artifacts view, which draws too much from the vocabulary of individual ethics.

Structural ethics studies social and material arrangements as well as components of such arrangements, such as artifacts and human agents. It has three aims: (1) to analyze the production of moral outcomes or consequences in existing arrangements and the role of different elements in this process; (2) to evaluate the moral goodness or appropriateness of existing arrangements and elements in them, and (3) to normatively prescribe morally desirable arrangements or restructurings of existing arrangements. In doing so, it also aims to identify, evaluate and prescribe roles of individual elements in these arrangements. Unlike individual ethics, structural ethics hence looks at larger structures and networks with the aim of engaging in social and technological engineering.

Let us consider an example of each of these three types of investigations. The first type can be illustrated with Latour's earlier example of the seat belt. The moral behavior of me wearing a seat belt can be analyzed as the result of not only my actions, but also the (inter)actions of other elements, including enforcement by the police and the behavior of my car, which in turn is the result of actions of engineers as well as safety advocates and policy makers. An analysis of this network of entities that influence my behavior can show how my moral behavior is shaped by this network, and it can assign a role to this effect to each of the entities.

The second type, aimed at evaluation, can be illustrated with cases in which a CCTV surveillance system in a public space is evaluated for its protection of the privacy of citizens. Such an evaluation requires that a whole network of human and nonhuman entities is being considered that play a role in safeguarding privacy. This includes evaluations of, amongst others, CCTV hardware and software, the properties and behaviors of the human operators, the protocols that govern their behavior, the characteristics of the room in which CCTV images are displayed or stored and their accessibility by third parties, and so on. All elements in this network, and their relations to each other, need to be evaluated relative to a set of privacy requirements, for their adequacy in safeguarding personal privacy.

The third type of investigation, aimed at prescription, can also be illustrated with reference to CCTV and privacy. This type of investigation would specify how a network surrounding a CCTV system would ideally be constituted so as to protect privacy and how its different elements would operate. Alternatively, recommendations could be developed for the improvement of an existing network, for example for the improvement of software, the training of operators, the improvement of facilities or procedures, and so on.

These three types of investigations focus on networks. However, they could also zoom in on particular components of these networks, including technological artifacts, and focus on their moral roles. For instance, it is possible to focus on the role of a particular CCTV software program in ensuring privacy within a particular network. It is also possible to consider this software abstracted from a particular network and consider its privacy-protective properties across a variety of possible networks. More generally, structural ethics can focus on both networks and components of networks, where these components can also be studied independently of a particular network. We may use the term *artifact ethics* for studies in structural ethics that focus on the moral role of technological artifacts in networks or across networks.

Artifact ethics, as a kind of structural ethics, has the advantage over moral agency approaches that it upholds important differences in the moral roles of artifacts and human agents, as discussed above. It moreover has the advantage of being able to attribute moral roles to artifacts, thus avoiding the fallacy that artifacts are morally neutral, while at the same time avoiding the false notion that morality can "reside" in artifacts, independently of their surroundings. In artifact ethics, it can be shown that artifacts sometimes constitute a major cause of morally good or bad consequences, while at the same time highlighting the dependency of these consequences on a larger network of things and humans.

Structural ethics requires a vocabulary to refer to the networks that are being studied as well as the different elements or component that these may contain, their relations to each other, and their behaviors. I will use the term "network" (or sometimes "arrangement" or "structure") to refer to structures of interacting entities that together determine outcomes or actions that are the subject of moral evaluation. The entities in networks include humans, artifacts, animals and natural objects, as well as larger structures composed of such entities. For instance, an organization is a larger structure that is composed of humans who work together towards a common goal, as well as nonhuman entities owned by the organization that are used to further this goal. An organization has itself a network structure, but it can also function as a component of a larger network in which it plays a role.

Relevant for structural ethics is the relative role of these different entities in fixing moral outcomes or behaviors. I propose that we call any entity in the network or arrangement that has a role in fixing moral outcomes or behaviors a *moral factor*. In ordinary English, a factor is an entity or component which contributes to an effect or result. This is the meaning I have in mind. At the same time, the word "factor" derives from the Latin *factor*, "who/which acts", and hence has associations with the notion of an agent. Moral factors shape or influence moral actions and outcomes. They have *moral influence*. The class of moral factors includes both human agents and various kinds of nonhuman entities.[2]

Moral factors can be positive or negative, measured against a moral rule or principle. A *positive moral factor* is one that contributes positively to a moral principle being upheld, whereas a *negative moral factor* contributes negatively. In addition, moral factors can be accidental or intentional. An *accidental moral factor* is one that happens to contribute towards a moral outcome in a particular arrangement. An *intentional moral factor* is one that has been intended to contribute to an outcome in a particular way. For instance, relative to the moral outcome of cars driving safely, a speed bump and a traffic controller would both be intentional moral factors, whereas a pothole that causes cars to drive slowly would be an accidental moral factor.

Whereas intentional moral factors are often positive, intentionally supporting moral principles, they can also be negative and intentionally contribute to violations of moral principles. For instance, relative to the principle of safe driving, a person imitating a police officer who maliciously signals drivers to perform unsafe maneuvers is an example of a human intentional negative moral factor. Oil intentionally spilled on a road is an example of a nonhuman intentional negative moral factor,

[2] Although moral factors are not necessarily bearers of moral responsibility, their role is moral in the sense that they have a role in generating moral outcomes. This role is a moral role and cannot be adequately captured by non-moral concepts like that of role responsibility. Role responsibility defines duties to others that are not necessarily moral. Nonhuman entities like artifacts cannot literally have duties (although they may have functions or roles), so they cannot be argued to have role responsibility. But even if they could, I would argue that independently from any role responsibilities they may have, they play (causal) roles in generating outcomes that can be evaluated as morally desirable or undesirable. It is these roles that require them to be included in ethical evaluation.

whereas oil accidentally spilled would be an accidental moral factor. If technological artifacts generate consequences that are positive or negative relative to a moral principle but were not intended by designers or users, then they are accidental moral factors relative to that principle.

Moral factors can be outcome-oriented or behavior-oriented. An *outcome-oriented moral factor* is a factor that contributes positively or negatively to the realization of a moral outcome. A moral outcome is a realized event or state-of-affairs that is the subject of moral evaluation. For instance, an unjust distribution of goods that results from an action or event is a moral outcome, as is harm to a person or a limitation to his or her freedom. Various moral factors can be identified as having caused these outcomes. A *behavior-oriented moral factor* is one that influences the moral behavior or actions of an agent. For example, my wearing a seat belt is a moral action that is influenced by various moral factors, such as blinking lights on my dashboard and police officers who check on seat belt use.

A structural ethics approach can account well for the distributed realization of moral norms in society, by showing that these norms are enacted not just by humans behaving according to them, but also by social and material structures being shaped to support these norms. A structural ethics approach can, as we have seen, account for the moral role of artifacts. It can also account for the role of things and humans in the moral behavior of human agents by identifying them as moral factors that are contributory causes of someone's moral behavior. Finally, a structural ethics approach can help solve the problem of distributed responsibility. This is the problem that when a moral outcome is the result of the actions of multiple agents, no single agent can be identified as being solely responsible for the moral outcome. A structural ethics approach can be used to analyze the role of different agents in producing the outcome, directly or indirectly. This analysis can then be used to assign moral responsibilities to these different agents. When technological artifacts are involved, it will only be human beings who are assigned responsibility, since technological artifacts and other items do not bear responsibility themselves, yet can serve as moral factors for which one or more human agents bear responsibility.

Individual ethics has a focus and aim that is different from those of structural ethics. Its focus is on the deliberations and actions of moral agents, instead of on networks or components of them. It has three aims that mirror the descriptive, evaluative and normative aims of structural ethics: (1) to study the moral principles, deliberations, traits and actions of human agents, (2) to evaluate the moral goodness of actions, judgments, and traits of human agents and attribute moral responsibility, and (3) to normatively prescribe what moral actions agents ought to perform, judgments they should hold, or traits they should have. Individual ethics makes use of the standard conception of a moral agent, and therefore concerns itself with human beings.

Structural and individual ethics differ in that they are concerned with the moral dimensions of different phenomena: networks and their components versus human beings and their actions. These phenomena require fundamentally different evaluations and subsequent interventions. In structural ethics, the primary aim of moral

evaluation is a better design of social and material arrangements: it is to investigate how components of networks can be rearranged, added or removed, through physical or social redesign, so as to generate better moral outcomes. Many networks have a public status or have public effects, so there is a public interest in their functioning, including their functioning according to public standards of morality.

In individual ethics, the objective of moral evaluation may likewise be to change individuals or their actions, but if so, this objective often does not translate into a plan for redesign but rather into a moral appeal to agents to change their behaviors or convictions. The emphasis on moral appeal in individual ethics stems from the fact that persons are generally believed to have free will and to bear the ultimate responsibility for their actions. In more extreme cases, agents may become the involuntary subject of "redesign", by means of involuntary therapy or treatment, or actions that are deemed immoral or harmful may be prevented through physical restraint or incarceration. In general, however, individual ethics is aimed at affecting moral deliberation in moral agents, whereas structural ethics aims to shape and redesign networks and their components.

Although structural ethics may focus on the moral role of particular components of networks, like artifacts, natural objects, and organizations, structural ethics does not focus on individual persons, since this is already the focus of individual ethics. In structural ethics, persons only appear as network components that have roles as moral factors relative to a nonhuman component which is the object of study or relative to the network as a whole. Individual ethics can be of service to structural ethics by improving its understanding of the moral role of particular humans in networks through an identification and analysis of their moral behaviors, values and beliefs.

Conversely, structural ethics can help individual ethics by identifying moral factors external to an agent that are relevant for the study or evaluation of his or her actions, traits or judgments. This is particularly helpful in moral explanation, moral evaluation and attributions of moral responsibility. Moral explanation may be improved by identifying the position of the agent in a larger network and the moral factors that contributed to an agent's action. For example, an explanation of why an agent committed a murder will be helped by an analysis of the material and social arrangements within which the agent was embedded, and the moral factors that directly contributed to his act, such as the availability of a gun and the presence of other agents who encouraged him. An evaluation of his actions and his responsibility for them may likewise take into account external moral factors that contributed to his action or detracted from it. In a similar way, external moral factors may have a role in the analysis or evaluation of moral beliefs, judgments, and traits.

## 8.6 Conclusion

I have argued that the moral artifacts view, according to which technological artifacts qualify as moral agents, brings us a better understanding of the moral role of technological artifacts, but at the same time obscures our understanding

of human moral agency by impoverishing the concept of a moral agent. I have proposed an alternative view, which I call structural ethics, which has the benefits of the moral artifacts view but also retains the standard conception of a moral agent, and thus the benefits of this conception. Structural ethics is supplementary to individual ethics, which focuses on (human) moral agents. It focuses on networks or structures consisting of human and nonhuman entities that have moral implications. I have called such entities moral factors. It can also account for the moral role of any kind of entity in producing moral behaviors and outcomes, including humans, animals, artifacts, natural objects, and complex structures like organizations. Artifact ethics is a division of structural ethics that focuses specifically on the moral role of artifacts, both within particular networks and across a range of possible networks.

So can no artifact ever qualify as a moral agent? Let us return to the moral intelligent agents view. Intelligent agents have a greater resemblance to human moral agents than any kind of artifact. They behave autonomously, they interact with an environment, they have a certain degree of intelligence and capacity for reasoning, they have intentional states of some sort, and they can be equipped with goals and moral categories and principles. So can they be moral agents? They could be if they meet the three criteria for moral agency that I outlined earlier. The first of these was a capacity for moral deliberation. Most intelligent agents do not have this capacity, so most would not qualify as moral agents. Progress is being made, however, to equip intelligent agents with capabilities for moral decision-making (Wallach and Allen 2008). So I will assume that some intelligent agents will be able to meet the first criterion (though see Johnson 2006 and Himma 2009).

The second criterion, being expected to adhere to standards of morality, is relatively easy to meet. It only requires that people expect intelligent agents not to act immorally. This expectation may already be in place, as we generally do not expect technological artifacts to be designed so as to produce unethical results. So I assume that intelligent agents can also meet the second criterion. The third criterion, that of moral responsibility and accountability, is the one that is hardest to meet. Most proponents of the moral intelligent agents view agree that it makes no sense to attribute moral responsibility to an artificial agent, and to praise or blame it for its actions. This is true both because artificial agents do not have free will and because they do not have the capacity to experience feelings like pleasure and pain (Johnson 2006; Himma 2009).

If this is true, there may be two ways to salvage the moral intelligent agents view. The first would be to redefine "moral agent" by dropping the moral responsibility requirement. This is what Floridi and Sanders (2004) propose. I find this solution unsatisfactory because moral agency has traditionally been identified strongly with moral responsibility. For intelligent agents who meet the first two criteria for moral agency but not the third, it would seem better to introduce a new term, such as "quasi-moral agent", which underwrites that these artificial agents are similar to, but in important ways different from, moral agents. A second way to salvage the moral intelligent agents view is by arguing that some intelligent agents can in fact be morally responsible. Intelligent agents can for example be programmed to

explain the moral deliberations behind their decisions, and to accept user input on the morality of their actions. This can be seen as a kind of responsibility or accountability. Still, as Stahl (2006) has argued, even such advanced systems lack some of the properties of full-blown moral agents, such as free will and a capacity to experience or feel blame and praise. He proposes that intelligent agents can at best have a sort of quasi-moral responsibility.

So it seems that some intelligent agents can significantly resemble moral agents, without fully qualifying as such. Such artificial agents, which may be called quasi-moral agents, have capacities for moral decision-making and possibly for responsibility or accountability through an ability to provide and receive feedback on their moral deliberations and actions. The vast majority of technological artifacts, however, including the vast majority of intelligent agents, do not qualify as moral agents, but do qualify as moral factors in the framework of structural ethics.

## References

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251–261.

Davidson, D. (1980). *Essays on actions and events*. Oxford: Oxford University Press.

Eshleman, A. (2009). Moral responsibility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2009 ed.). URL: http://plato.stanford.edu/archives/win2009/entries/moral-responsibility/

Floridi, L., & Sanders, J. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Himma, K. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*(1), 19–29.

Johnson, D. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*, 195–204.

Johnson, D., & Powers, T. (2008). Computers as surrogate agents. In M. J. van den Hoven & J. Weckert (Eds.), *Information technology and moral philosophy*. Cambridge: Cambridge University Press.

Jonas, H. (1984). *The imperative of responsibility: In search of ethics for the technological age* (H. Jonas & D. Herr, Trans.). Chicago: University of Chicago Press.

Keulartz, J., Korthals, M., Schermer, M., & Swierstra, T. (2004). Pragmatism in progress: A reply to Radder, Colapietro and Pitt. *Techné: Research in Philosophy and Technology, 7*(3), 38–48.

Latour, B. (1987). *Science in action*. Cambridge, MA: Harvard University Press.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change*. Cambridge: MIT Press.

Powers, T., & Johnson, D. (2004). *The moral agency of technology*. Paper presented at the 2004 workshop on understanding new directions in ethics and technology, University of Virginia. Unpublished, 28 pp. Available online at http://www.sts.virginia.edu/E&T2004/pdf/MAT.pdf

Searle, J. (1984). Intentionality and its place in nature. *Synthese, 61*(1), 3–16.

Stahl, B. (2004). Information, ethics, and computers: The problem of autonomous moral agents. *Minds and Machines, 14*(1), 67–83.

Stahl, B. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology, 8*, 205–213.

Sullins, J. (2006). When is a robot a moral agent? *International Journal of Information Ethics, 6*, 12. Retrieved at http://www.i-r-i-e.net/issue6.htm

Verbeek, P. P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. University Park: Penn State University Press.

Verbeek, P. P. (2008). Obstetric ultrasound and the technological mediation of morality – A post-phenomenological analysis. *Human Studies, 31*(1), 11–26.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics, 24*, 227–248.

# Chapter 9
# Artefactual Agency and Artefactual Moral Agency

**Deborah G. Johnson and Merel Noorman**

**Abstract** This chapter takes as its starting place that artefacts, in combination with humans, constitute human action and social practices, including moral actions and practices. Our concern is with what is regarded as a moral agent in these actions and practices. Ideas about artefactual ontology, artefactual agency, and artefactual moral agency are intertwined. Discourse on artefactual agency and artefactual moral agency seems to draw on three different conceptions of agency. The first has to do with the causal efficacy of artefacts in the production of events and states of affairs. The second can be thought of as acting for or on behalf of another entity; agents are those who perform tasks for others and/or represent others. The third conception of agency has to do with autonomy and is often used to ground discourse on morality and what it means to be human. The casual efficacy and acting for conceptions of agency are used to ground intelligible accounts of artefactual moral agency. Accounts of artefactual moral agency that draw on the autonomy conception of agency, however, are problematic when they use an analogy between human moral autonomy and some aspect of artefacts as the basis for attributing to artefacts the status associated with moral autonomy.

D.G. Johnson (✉)
Department of Engineering and Society,
University of Virginia, Charlottesville, VA, USA
e-mail: dgj7p@virginia.edu

M. Noorman
eHumanities Group, Royal Netherlands Academy
of Arts and Sciences, Amsterdam, The Netherlands
e-mail: Merelnoorman@gmail.com

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_9,

## 9.1 Introduction

This chapter takes as its starting place that artefacts, in combination with humans, constitute human action and social practices, including moral actions and practices.[1] Identifying and differentiating the entities that make up the world is the work of ontology, and the ontology implicit in ordinary language and informal thought seems to presume three fundamental kinds of entities: natural, human, and artefactual. Artefacts are individuated as entities through mental acts that separate human-fashioned materiality from naturally occurring materiality and from human activity and meaning. This ontology is the backdrop against which questions of agency typically arise. That is, having divided the world into categories of things, scholars and theorists ask where agency is to be found. Generally, humans are presumed to have agency, while the agency of nature and artefacts are in dispute (each in distinctive ways). And, once the question of agency is raised, the further question of moral agency comes into focus. If artefacts have agency, why would they not have moral agency?

Ideas about artefactual ontology, artefactual agency, and artefactual moral agency are intertwined. In order to get a handle on the debate about artefactual moral agency, artefactual ontology and artefactual agency must first be addressed. After making the case for artefacts to be understood as components in larger sociotechnical systems, we distinguish three conceptions of agency: causal efficacy, acting for, and moral autonomy. We then take up the issue of artefactual moral agency arguing that conceiving of artefacts as moral agents can be productive when it refers to the causal efficacy of artefacts or to the tasks that have been delegated to artefacts by humans. However, understanding artefactual moral agency in terms of moral autonomy is problematic.

## 9.2 Artefactual Ontology

Artefacts are defined and generally understood to be human-made material objects. Although, as already suggested, the ontology embedded in our language and ways of thinking and speaking presumes three kinds of entities, when pressed, most of us acknowledge that the things in these categories overlap. We make statements of the following type: ‘humans are part of nature’; ‘artefacts are made by humans’; ‘nature constrains what humans can do’; and ‘artefacts are made by manipulating nature.’ So, although the three types of entities are distinguished, they are inseparable; they are incomprehensible separately. Artefacts do not exist without humans making them; humans are part of the natural world; nature is understood as the ‘stuff’ from which humans come. This inseparability means that artefacts are never just artefacts.

[1] This material is based upon work supported by the National Science Foundation under Grant No. 1058457. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Particular artefacts are individuated as entities by mental acts that draw ontological lines. To comprehend the significance of line drawing, consider the refrigerator of one of the authors.[2] Deborah's refrigerator is an artefact, that is, the chunk of plastic and metal that sits in her kitchen is an artefact. Some might even say that her refrigerator is an autonomous entity (artefact) because it maintains its internal temperature "on its own". The thermostat in her refrigerator detects the temperature and signals other components of the refrigerator to change states so as to raise or lower the internal temperature. In this respect Deborah's refrigerator might be described as an autonomous agent acting on her behalf. Admittedly, her refrigerator's so-called agency does make a difference in her life. It allows her to conveniently eat and drink all kinds of things that might otherwise spoil.

The problem with characterizing Deborah's refrigerator as an artefact (and especially as an autonomous artefact) is that it only keeps her food cool when it is plugged into an enormously complicated power grid. Indeed, her refrigerator can easily be understood not to be an entity in itself but to be (merely) a component in a larger *technological system*. It is connected to a complex of artefacts – the electrical socket in her kitchen, the wires that run through her house and out to the street, the power station maintained by a company named Dominion Virginia Power. Going the other way, that is, breaking the rectangular chunk of metal and plastic into its component parts also suggests a technological system, for the rectangular chunk of materiality sitting in Deborah's kitchen is itself a combination of many different artefacts – a motor, vents, wires, metal parts, plastic shelves, etc.

The fact that Deborah's refrigerator is a technological system (meaning that it is multiple chunks of metal and plastic) and that it is a component in a larger technological system, is, however, only part of the story. A refrigerator only works as a refrigerator when human beings behave in certain ways. Deborah has to plug the rectangular chunk of metal and plastic into a socket; an electrician had to lay wire connecting the socket to a power grid; all the people working for Dominion Virginia Power have to come to work each day and do their jobs. In fact, the institutional arrangements constituting the power station are an enormous feat of human social organization and cooperation. In addition to those who work at Dominion Virginia Power are many other human beings and especially Deborah. She is needed to buy food, to open and close the door to put the food in and take it out. Importantly, she has to pay her utility bill (or else the Dominion Power will disconnect her refrigerator from the grid). Other humans are involved as well, for she could not buy food that needs refrigeration unless grocery stores carry it, and this in turn requires that trucks and airplanes bring refrigerated foods from far off lands to her grocery store. So, Deborah's refrigerator is not just a technological system; it is a *sociotechnical* system.

Where does the entity that Deborah calls her refrigerator begin and end? It seems that we have collectively and conventionally drawn a line. We have

[2] To be sure, refrigerators are more complex than, say, forks and bowls or hammers, but a complete typology of artefacts would take too long to introduce here. Later the distinction between computational and non-computational artefacts will be addressed.

decided we will count the rectangular chunk of plastic and metal (the artefact) that sits in her kitchen as 'a refrigerator.' We have decided to leave on the other side of the line (outside of the concept of refrigerator) such components as the electrical grid to which her refrigerator must be connected, all the people who maintain the electrical grid, Deborah who must open and close the door to put in and take out food, the grocery stores, the trucks that deliver items to the grocery store, the global trade markets that bring foods needing refrigeration to her grocery store, and so on.

The ontological line that we draw delineates Deborah's refrigerator as an artefact. In doing this we mentally and selectively extract it from the world in which it functions and has meaning; we disconnect it from all the other entities (human and material). In doing so, we make it 'something'; we think of it as something in itself. The mental act of thinking of it as an artefact blinds us to all of the activity behind the scenes (offstage), activity that makes her refrigerator function in the way she has come to expect. That her refrigerator is a sociotechnical system, that it achieves its results through a combination of human and non-human activity becomes something that we must work to see, against the backdrop of the artefactual ontology implied in ordinary language. It is not that we can never understand the connections among parts; obviously we can. The point is that the ontology draws attention to some of what is going on and directs attention away from other things that are going on.

Some might say that what we have just explained is the difference between artefacts and technology. Artefacts are material objects; technologies are sociotechnical systems. Artefacts are components in sociotechnical systems. This framework would seem to allow us, then, to ask what part artefacts play in sociotechnical systems and to ask whether the artefacts have agency.

Some of those who are particularly focused on computational artefacts might accept the distinction between artefacts and technology but insist that computational artefacts are different – because they are more autonomous or because they are autonomous in a distinctive way. This difference might then mean that computational artefacts can have agency when other artefacts do not. Consider an automatic pilot system. We can easily think of an automatic pilot software system as an agent acting on behalf of humans. It does many of the tasks that human pilots used to do and still do (when the automatic pilot is turned off). Of course, the reason the automatic pilot can control the airplane is because it was designed to do so and has been delicately connected to various other components of the airplane. So, whether or not the automatic pilot is autonomous is not a simple matter and whether its agency is different from other artefacts is not obvious.

Automatic pilots function only in combination with humans. In the design of automatic pilot systems, humans decide when and how the automatic pilot takes control of the airplane. Indeed, how automatic pilots work with humans can vary. Most automatic pilots are designed so that they take over control of the plane only when humans tell them to (e.g., when a human flips a switch). Of course, they could be designed so that they take control independent of any immediate human activity (that is, when they receive signals from other artefacts, internal to the airplane) or

when humans at remote locations do something. And, of course, humans could decide to assign more and more of this decision making (i.e., when to go into automatic pilot control) to the technological components. The point, however, is that the automatic pilot, like the human pilot, *only* functions when it is a component in a larger sociotechnical system. To refer to the automatic pilot as an agent is then to draw a line around a particular part of that system. One might do this in order to draw attention to its behaviour or its significance apart from or perhaps in interaction with the other components in the system.

Much of this is well-trodden territory. STS scholars have been especially concerned with how and why lines (boundaries) are drawn between humans and machines. For example, Suchman (2001) writes: "I take the boundaries between persons and machines to be discursively rather than naturally effected, and to be always available for refiguring." That lines are drawn between humans and machines (or artefacts) goes hand-in-hand with lines being drawn around artefacts.

The lines drawn are not innocent, they have real social and material consequences (Barad 1996). Lines are drawn to make sense of the world, to facilitate practices, to give meaning, to achieve tasks. Delineating 'refrigerator' as an artefact containing shelves, doors, freezing elements, wires, nuts and bolts, rather than as a sociotechnical system, may make it easier to talk about a particular part that can be pointed to, moved, chosen, sold, etc. Yet, alternative ontologies are possible and can make a difference in what is seen and understood. For example, in Chap. 3 of this volume, Introna argues for a new ontology that better reflects the co-constitution of artefacts and humans. He emphasizes how each part is what it is because of other parts and traditional (human-artefact) line drawing works against our being able to notice this.

## 9.3 Artefactual Agency

Given the intricacies of delineating artefacts, what does it mean to say that artefacts have agency? It seems odd, on the face of it, that the question of agency would be raised with respect to 'things' that have been mentally constructed as chunks of materiality. Why draw lines around a chunk of materiality, extracting 'it' from a dynamic socio-material whole, and then ask whether (or proclaim that) the delineated chunk has agency?

One plausible answer to this question is that attributing agency to artefacts draws attention to (emphasizes, punctuates, makes visible) the role and significance of chunks of human-fashioned materiality in constituting the human world. This would, in turn, draw attention to the importance of decisions about fashioning and deploying those chunks of materiality. Another plausible answer (not unrelated to the first) is that thinking about artefacts as having agency is a useful way of understanding those chunks; it allows us to see aspects of materiality that we might not otherwise notice. For example, thinking of artefacts as having agency might allow us to see that they are far from inert, passive or neutral.

Although both answers seem plausible, more seems to be at stake in the discourse around artefactual agency. Attributions of agency to artefacts seem to do more than claim that agency is a useful concept. Indeed, since humans use language in complex, creative, and often fanciful ways, attributions of agency to artefacts may have a variety of functions or meanings or illocutionary uses. This is all the more likely because agency is such an unclear concept (Lee and Brown 1994). Agency generally refers to the ability or capacity of an entity to act in the world. However, as explained below, many different conceptions of agency have been articulated and used in particular contexts.

Ironically, although unclear, agency is an important concept. It anchors many important discourses – moral discourse; discourse about what it means to be human; discourses about human relationships with animals, the earth, and transcendental beings; discourse about human rights and discourses about power and accountability. Indeed, the fact that agency is such an important concept and that it is so poorly understood may be connected; that is, its blurry meaning may facilitate use of the concept of agency in so many different contexts.

## 9.4 Three Conceptions of Agency

Discourse on artefactual agency seems to draw on at least three different conceptions of agency. The first has to do with causality. Many attributions of agency point to the *causal efficacy* of artefacts in the production of events and states of affairs. The second conception of agency might be thought of as *acting for* or on behalf of another entity: agents are those who perform tasks for others and/or represent others. The third conception of agency has to do with *autonomy*; agents are entities with the ability to think, decide, and intend, and to act accordingly. Distinguishing these three conceptions of agency is key to understanding discourse about artefactual agency and artefactual moral agency. Problems arise when one conception is conflated with another.

## 9.5 Causal Efficacy

If one wants to explain how the world got to be the way it is or how it currently works, or if one wants to shape the world of the future, thinking about causality seems unavoidable. One need not be a determinist to accept that things happen because of things that came before; one does not have to be a determinist to recognize that to get to a future state, events and changes will have to occur between now and then. Much of the discussion of artefactual agency – explicitly or implicitly – seems to have to do with the causal efficacy of artefacts in bringing about states of affairs. The design and availability of artefacts facilitate and constrain human behaviour. Whether the artefacts operate independently in time and space from

humans (as in the case of thermostats controlling the temperatures in refrigerators) or they are deployed via direct human control (as when a person presses the trigger on the gun), artefacts make a difference in what humans do and what happens in the world. The availability and design of particular artefacts affects how humans think, act, and organize themselves.

Causal efficacy is at least part of what is claimed by many STS theorists when they refer to the agency of artefacts. STS scholars have emphasized the affordances and constraints of artefacts and the contributions they make to outcomes, i.e., ongoing states of affairs, predicted futures. Actor Network Theory (ANT) is a good case in point (Law and Hassard 1999); in treating nature, artefacts, and humans symmetrically, ANT acknowledges the causal contribution of all three in technological outcomes. Of course, ANT theorists insist – in effect – that the causal efficacy of each node in a network is dependent on other nodes. For this reason, it may be more accurate to say that ANT draws on the notion of causal efficacy, but is not reducible to it. Notice that in being ecumenical about artefacts, humans, and nature, ANT denies any special (a priori) status for any category of entity.

Although artefacts have causal efficacy, it is important to remember that artefacts only have causal efficacy in combination with humans. Humans design and deploy artefacts. Artefacts have meaning and function in relation to humans and human endeavours. We can draw lines around artefacts and we can extend or interpret the concept of agency so that we think and speak of artefacts as acting, but in doing so, we run the risk of pushing out of sight the human activity that is intertwined with the non-human activity. Suchman refers to this human activity as the 'offstage' or 'behind the scenes' activity that we may not notice but without which machines/artefacts do nothing (1998). Remember the refrigerator can only keep Deborah's food fresh if she plugs it in, pays her electricity bill, buys the food at a grocery store, and only if Dominion Virginia employees come to work every day, etc. All of this activity may be invisible when we think of the refrigerator as an agent.

## 9.6 Acting For

Human activity is explicit in the second conception of agency. Discourse on the agency of artefacts often seems to involve the idea that agents are those who perform tasks for humans or act on behalf of humans. In legal contexts, agents are those who are authorized to negotiate on behalf of a principal, as in the case of real estate agents or literary agents (Heath 2009). Here agency involves representation, though the representation involves the agent using his or her expertise to perform tasks for the client. Latour's analysis of artefacts in "Where are the missing masses?" (1992) draws on this type of agency together with causal efficacy. He treats artefacts as if their role is to replace human actors; that is, artefacts do the (causally efficacious) work that human actors used to do or would have to do were the artefacts not there. He describes this as machines being delegated part of *the program of action*. The

mechanical door groomer replaces the human that stood in front of the door and opened and closed it as people came along; the traffic light replaces the police officer standing in traffic and using hand signals. These artefacts perform delegated tasks both on behalf of those who deployed (situated) them and those who encounter them. Regulators and engineers placed the traffic light at a crossroad to act on their behalf in enforcing moral behaviour from drivers, cyclists and pedestrians. They delegate this task or act by inscribing their intentions in the design of the traffic light; the traffic light expresses these intentions in signalling people when to stop and go. Similarly the mechanical door groomer was situated by architects, builders, and building owners to direct people to a particular place and to assist them in entering and leaving the building. It is thus not just the causal efficacy that is important here; it is the idea that artefacts affect human action through the delegated intentions inscribed in their design.

In a similar way, some computer scientists use an 'acting for' conception of agency to describe interactive software programs that accomplish tasks on behalf of their users, such as finding relevant news items online. Such *artificial agents* may perform decision-making tasks as well as negotiate with other agents. Computer scientists use the term agent here to mark a difference with other kinds of computer technologies. That is, artificial agent programs are different in that they are able to learn their users' interests, habits and preferences and use this information as they roam the Internet and carry out tasks for the users.

The 'acting for' conception of agency draws on a metaphor. Calling the mechanical part of the doorframe a door groomer makes an analogy with the human door groomer; calling an autonomous vacuum cleaner a housekeeper makes an analogy with the human housekeeper. Similarly, computer scientists and others refer to computer programs as software agents *as if* they acted on our behalf in the way that a servant or a hired worker might.[3]

Metaphors are more than ornamental devices or tools of persuasion in rhetoric. They are useful in making the unfamiliar, familiar. Metaphors help us to understand and make us comfortable with what might otherwise seem too complicated or alien. They allow us to see aspects of a thing that might otherwise be opaque. For example, referring to certain kinds of software as software agents that work on our behalf may help us to understand and explain what a complex piece of computer code is intended to do and how it is supposed to relate to the human user. Thinking about software programs *as if* they were agents explains and helps in understanding and designing computational artefacts (Noorman 2009). Similarly, describing machines as delegates that substitute for human actors helps to draw attention to the role they perform in shaping human actions and morality.

Metaphors, however, are not innocent; they can sometimes even be dangerous. They draw attention to particular similarities between two things, using one that is presumably well understood, to help understand one that is not. However, in thinking metaphorically, we may be directed to think that the two things have more in

[3] Johnson and Powers (2008) used the metaphor of computer systems as surrogate agents to tease out the possibility of a form of responsibility for computer systems.

common than they do. Important and relevant dissimilarities between the compared entities may be pushed to the background by making a particular analogy between the two entities. Moreover, analogies can lead us to believe we understand something when in fact the thing used in the analogy is very poorly understood, e.g., human consciousness. We have to be careful, then, in drawing analogies between relationships in which humans 'act for' other humans and relationships in which artefacts 'act for' humans. For example, income tax preparation software may be thought of as your personal tax accountant (agent), but software and a human accountant are different in important ways.

## 9.7 Moral Autonomy

A third conception of agency involves autonomy. Traditionally, autonomy was thought to be a distinctive feature of humans differentiating them from other kinds of entities. Because they have autonomy we think of humans as having agency. Humans think, choose, decide and then act. Humans act for reasons and their intentional behaviour is outside the ordinary realm of material causality. On the other hand, artefacts do not act for reasons; their behaviour is the result of causality, be it deterministic or non-deterministic.

Human autonomy is what makes morality possible. That is, morality applies to humans and not to animals and machines because humans have autonomy. In moral theory 'ought implies can'. If a being does not have the capacity to freely choose to act (autonomy), then it does not make sense to have a system of moral rules specifying what that being should do. This idea is famously captured in Kant's distinction between things that behave according to natural law and things that behave according to the conception of law. So the autonomy conception of agency is intertwined with a set of ideas about human capacities, action, intentions, and differences between humans and other kinds of beings.

To be sure, this conception of autonomy continues to perplex moral philosophers and many others. Philosophers and ethicists continue to try to explain how it is possible (and whether it is true to say) that humans are free and have consciousness and autonomy. Whether autonomy and consciousness are amenable to reductionist accounts is an issue that will not be taken up here.

Autonomy is also used in other, non-moral contexts to describe artefacts that operate independently from humans. Remember Deborah's refrigerator was thought of as autonomous because it maintained a particular internal temperature without any action on Deborah's part. Similarly, we speak of autonomous vehicles, autonomous systems, autonomous robots, etc. Computer scientists refer to particular programs as autonomous in order to highlight their ability to carry out tasks on behalf of the user and to perform those tasks independently. As a result of machine learning algorithms, for instance, these programs are thought to be more capable of operating independently in unknown environments than pre-programmed computers systems.

Thus, the autonomy conception of agency seems to include two different conceptions. One has roots in the notion of human moral autonomy and the other refers to the independence of things from immediate control by humans. These two different ideas should not be conflated, for human moral autonomy provides a foundation for establishing moral status or moral standing. Because humans are autonomous beings that can choose to act in the world, certain rights can be attributed to them and they can be held responsible for their actions. The other conception of autonomy has little to do with morality or moral standing. It is agency only in the sense that it identifies something as operating independently. Having the capacity to operate independently is not sufficient to justify moral status or standing. For this reason, the autonomy conception of agency has to be used cautiously.

## 9.8 Artefactual Moral Agency

All three conceptions of agency are found in debates about artefactual moral agency, though some authors rely more heavily on one conception or another and some combine or conflate several conceptions. The different conceptions are used to clarify aspects of the role of artefacts in morality. Keeping the three conceptions in mind should facilitate discussion of artefactual moral agency; failure to distinguish them runs the risk, among other things, of overlooking important asymmetries between humans and artefacts. The causal efficacy and acting for conceptions are particularly important in understanding moral consequences but neither has implications for the moral standing of artefacts. The autonomy conception of agency, on the other hand, provides a foundation for moral standing.

The causal efficacy of artefacts is generally the foundation of claims about artefactual moral agency. For example, in Chap. 5 of this volume, Verbeek emphasizes the role of artefacts as mediators. Although Verbeek does not explicitly use the language of causality, his account shows how artefacts affect human experience. Verbeek does not want his account to be interpreted as a claim that in mediating human experience, artefacts entirely determine what happens. This concern belies a causal notion at work in his thinking. In his account, artefacts have an active, but not a final, role in organizing relations between humans and world. In order to better understand this role, he pushes for a distributed or a ‘composite’ conception of moral agency: agency is not an inherent property of either humans or artefacts; it is the outcome of the interactions between humans and things. He explains that “in their own way – distinct, but not separated – humans and things contribute to moral actions and decisions”. ‘Contribution’ seems here very close to, if not the same as, causal efficacy.

Because of their casual efficacy, artefacts make a moral difference. They make a difference in moral practices, moral outcomes, and even moral notions. Think about the difference in the kind and degree of privacy that individuals have now as so many activities have been configured or reconfigured around computers and information technology. Consider changes in the nature of familial relationships

accompanying the use of cell phones, reproductive technologies, and child-rearing devices. Artefacts make a moral difference both for better and worse. A world with aqueducts, bridges, antiseptics, sanitation systems, and bicycles is a world in which humans have more pleasant lives and may live longer. Artefacts facilitate individual moral practices, e.g., playpens help parents keep their children safe and ambulances bring medical treatment quickly to those who need it. Of course, artefacts also work in the other way: landmines help to kill and maim the innocent; electronic devices are used to intrude on personal privacy; and so on. Artefacts affect how we fulfil obligations, keep promises, distribute resources, etc.

So in considering the active role of artefacts in moral actions and practices, we can meaningfully think of artefacts as moral agents by treating agency as causal efficacy in the production of states of affairs or events that have moral consequences. Artefactual moral agency means here simply that artefacts have a role in moral actions and outcomes; they affect or make a difference in moral actions and outcomes.

The 'acting for' notion of agency also grounds a conception of artefactual moral agency, that is, 'acting for' gets at something that is important about the role of artefacts in morality. Artefacts can be said to be moral agents in the sense that (or when) they are delegated tasks that are either constitutive of moral practices or have moral consequences. As Latour suggests, we delegate tasks to artefacts to achieve certain results and when we do so, we effectively treat artefacts as our agents. When artefacts perform delegated tasks that constitute states of affairs with moral features or moral consequences, the artefacts can be thought of as our moral agents. When a hospital machine keeps a person breathing, we might think of the machine as a moral agent; when a cell phone allows parents to keep track of their children, we might think of the cell phone as an assistant in fulfilling parental duties. These artefacts perform delegated tasks that constitute moral practices and have moral significance. In the same way, a landmine might be thought of as an immoral agent in the sense that it has been delegated the task of killing or maiming those who step on it. The landmine acts on behalf of those who have intentionally put it in a particular location.

Remember, however, that the acting for conception of artefactual moral agency is metaphorical and, as mentioned earlier, we have to be careful in using metaphors. Thinking of artefacts as if they are agents that perform tasks on behalf of human actors helps to understand how artefacts can shape moral action. Designers and engineers seem to delegate morality to these artefacts by inscribing their intentions in the design of the artefact in order to affect the user's actions in morally significant ways. However, it is a step too far to claim that humans and artefacts are interchangeable components in moral action. Though they perform some similar tasks, the traffic light and the police officer directing traffic are not morally the same.

One important difference between humans acting as agents for others and artefacts acting as agents for humans is responsibility. Central to the acting for conception is the idea of delegating tasks; we delegate tasks to humans and to artefacts. Importantly, however, human-to-human delegations generally involve tasks *and* responsibility; human-to-artefact delegations involve tasks but no responsibility. The hospital machine and the landmine perform actions on someone's behalf but neither is (or is

considered) responsible (except in a causal sense) for the outcome. We might identify the artefact as the point of failure when we do not get what we expected, but we typically hold the humans who produced or deployed the artefact to blame. Those who design and maintain the artefacts are typically considered responsible both for the accomplishments and failures of artefacts.

There is, thus, a significant and not to be overlooked difference between human-to-artefact and human-to-human delegation relationships. Humans delegate tasks with responsibility to human agents recognizing that the delegate has the capacity to negotiate and re-negotiate with them about appropriate goals and strategies. On the basis of this and the agent's expertise and, often, experience, clients delegate a range of decision-making latitude to the agent and the authority to use it. A literary or press agent is authorized to act on a client's behalf with regard to a particular situation or set of decisions and transactions, e.g., finding a publisher and obtaining a contract for a particular book. The agent may, for example, be authorized to negotiate with others and constrained to develop only the preliminary terms of a contract. Nevertheless, the required actions for these negotiations cannot be fully designated at the beginning or specified in rigid rules or static models of behaviour. The literary agent is authorized to behave as she thinks appropriate in contract negotiations, in order to achieve the desired outcome. She will make judgments in contingent situations drawing on her expertise and background knowledge, and her understanding of her client's wishes.[4] She is responsible for these decisions, and can be called upon to account for and explain her decisions and actions, which may result in praise or blame.

Responsibility is not part of human-to-artefact delegations, that is, when humans delegate tasks to artefacts, they do not delegate responsibility to the artefact. In the delegation of tasks to an artefact, the client defines and redefines, distributes and redistributes tasks so that the artefacts behave according to well-defined rules and protocols (Collins and Kusch 1998). Such delegations may allow variability in artefactual behaviour, but only if humans are indifferent to the variability. For example, the speed with which the hands of the clock move might vary by milliseconds, but the humans who look at the clock do not notice or care. On the other hand, if the artefact behaves in unexpected ways or counter to its delegated task, the artefact is considered flawed; it is not considered irresponsible. A clock can spin its hands clockwise, but when the hands spin backwards, we say it *malfunctions*.

[4] Because human agents 'acting for' have decision-making latitude and the possibility of renegotiation, trust is an important aspect of human-to-human delegations. Clients must trust that their agents will use their decision-making latitude in accordance with specified constraints. Successful delegation relationships generally build trust, that is, the more an agent acts successfully on a client's behalf, the more the client is likely to trust the agent in the future. Arguably, trust is involved in person-to-artefact delegations. We trust refrigerators to keep our food cold, search engines to bring us relevant results, etc. We speak not just of trustworthy accountants but also of trustworthy (reliable) computing. Of course, what it means to trust a human agent to act on one's behalf and what it means to trust an artefact or a technological system to act on one's behalf are quite different.

We do not have meaningful practices of blaming clocks or holding them responsible. We might, in fun, say that the clock is behaving badly (stretching the agency metaphor), but we think it is broken. We blame the humans who made the clock. Even were we to have some practice of "blaming clocks" the meaning of saying they are responsible would be blurry at best, and more likely incoherent. When it comes to artefacts, responsibility is traced back to human operators, designers, managers, or even politicians: there is a bug that needs to be fixed, developers or users did not have enough training, or the conditions in which the artefact would be used were not accurately anticipated.

The *acting for* conception of agency, thus, provides another meaningful way to think about artefactual moral agency, but only in a narrow sense. Artefacts have moral agency in the sense that they are delegated tasks that constitute moral practices and have moral consequences. This conception of artefactual moral agency draws attention to certain, previously overlooked, aspects of the role of artefacts, but it does not provide a basis to blur the boundaries between humans and artefacts. The metaphor only goes so far.

Failure to recognize the metaphorical character of accounts describing artefacts as agents acting for humans may lead to conflating this conception of agency with the human autonomy conception of agency. This is especially problematic because the autonomy conception is embedded in a set of ideas that refer to and elucidate the capacity for responsibility. It is tied to what it means to be human. Artefacts do not have the kind of autonomy that has traditionally, and non-reductively, been associated with bearing responsibility for one's actions. Some may argue that it is possible to reduce responsibility to something that applies to artefacts, but to do so would seem to violate something very fundamental about the conception. On the other hand, some may argue that the autonomy conception of agency is *the* only conception of moral agency, i.e., only if artefacts have autonomy can they be considered moral agents. As shown above, this position goes too far and fails to recognize the important role of artefacts in morality.

The autonomy conception of agency is important not just because it grounds morality, but because it confers a particular kind of status. Historically, the autonomy of humans is what distinguished humans from other animals in the chain of being. One reason for distinguishing between humans, animals, and machines is to identify which entities should be accorded rights, especially rights against entities in other categories. Differences in status affect, for example, the right to own property, to vote, to have freedom of speech, to have privacy and, in daily life to determine one's own actions. Differences in status affect negative and positive rights; they lead to ideas about which entities must refrain from killing or keeping captive which other entities. In the Christian-Judaic moral tradition, entities that have autonomy have a special status. They are accorded rights though they are also assigned responsibility.

In Chap. 8 of this volume, Brey argues against blurring the distinction between humans and artefacts on grounds that doing so diminishes the moral status of humans. He gives two reasons for this. First: "the classical notion of an agent has an important role in our moral image of a human being." Brey argues that when

artefacts are called agents, the "special features of human agency are lost" and "the moral image of humans is damaged as a result." Brey's second reason has to do with explaining and accounting for events. He affirms the distinction between explaining events and explaining actions wherein the latter involves reasons and intentions and the former involves causality. He argues that extending the notion of agency to include artefacts will destroy the distinction between actions and events, and eliminate "the special role of actions in our understanding of the world". By embracing a conception of agency that requires autonomy, he draws an ontological line and thereby reserves a special status for human actors, as being the only entities capable of acting. The consequence of this is that humans can still be the centre of morality.

Extending the autonomy conception of agency to artefacts has implications both for diminishing the moral status of humans as well as for increasing the status of artefacts. Implications of the latter kind are perhaps most salient in the discourse on computational artefacts. In the context of computer science, as mentioned before, researchers sometimes use the concept of autonomy to describe how computer systems are capable of performing particular tasks independently. However, when the autonomy conception of agency is used to challenge ontological lines between humans and computers, the conception becomes problematic. Take for instance the discussion that focuses on the idea that human behaviour and cognition can be understood in terms of computational processes. If the autonomy of human beings can be analysed and explained as a series of computations, then it can be formalized and simulated by computers. This discourse is mostly speculative with philosophers and cognitive scientists imagining and speculating that computation may someday produce entities that are not just autonomous but have the capacity for moral autonomy. Some speculate that this will necessitate the granting of rights to these computational entities. Although this will produce only a simulation of human moral autonomy, those who believe deeply in the computational model seem to believe that the equivalence between computational moral autonomy and human moral autonomy would justify the attribution of moral agency to these autonomous computational entities.

Although the suggestion that machines might have moral autonomy seems misconceived, the status of things and humans is not immutable. As mentioned earlier, ANT treats human, natural and artefactual nodes symmetrically in analysis, in order to discover how they are constructed. In fact ANT does not assume that there are three types of entities *a priori*. Rather the categories, natural, human, and artefactual, are the outcome of negotiations, not a given. From this perspective, the difference in status between humans and artefacts is historically and culturally constituted (Suchman 1998); humans have been constituted as unique, as the ultimate reference point. This would suggest that status can shift and change. However, such changes would not be without broader consequences to, for instance, social practices in ascribing responsibility or attributing rights.

Using the autonomy conception of agency in relation to artefactual moral agents is problematic in the sense that we are asked to imagine or hypothesize that machines will at some point in their development operate in a way that

would justify considering them morally autonomous, ascribing responsibility to them and granting them the status of moral agents. Without knowing how such machines would work, this idea seems to go from using a metaphor to understand certain phenomena, to using it as a basis to attribute status. To be sure, there are and are likely to be in the future, similarities between computational machine autonomy and human autonomy. This is not surprising since humans will build such computational machines to accomplish humanly conceived tasks. However, attributing artefacts a moral status comparable to humans would affect the instrumental status that artefacts now have. Human-artefact relationships have many non-instrumental dimensions, but because of their status, artefacts can be treated merely as means. They are not expected to make judgements based on their own motivations or desires (even if that were possible), nor are they expected to account for their decision-making in the way that humans are. Humans define the boundaries of acceptable behaviour for artefacts, and allow them to operate within these constraints. By moving from metaphor to status these complementary statuses would be compromised.

## 9.9 Conclusion

Our analysis suggests, then, that attributions of moral agency to artefacts make sense when they refer to the causal efficacy of artefacts and when they refer to the tasks that have been delegated to artefacts by humans. There may well be other intelligible accounts of artefactual moral agency, but we have argued that using autonomy as the basis for artefactual *moral* agency is problematic. Attempts to extend moral autonomy to artefacts seem to move from a metaphor to a claim of moral status, that is, they claim humans and machines are analogous and, then, on the basis of the analogy attribute to artefacts the status (or potential to have the status) associated with moral autonomy.

However, there are good reasons for keeping the status of humans and artefacts different, that is, for keeping humans and artefacts in different categories with regard to agency. For instance, in separated categories they can be treated as complementary rather than equivalent. More importantly, in order to maintain human responsibility for the development and deployment of artefacts, an anthropocentric moral perspective is essential (Johnson and Miller 2008). Whether we justify attributions of moral autonomy and responsibility on utilitarian or non-utilitarian grounds, attributions of moral responsibility (the expectation that one will be held to account) have the effect of shaping human behaviour. And if one wants to shape the behaviour of artefacts, then we ought to hold onto practices that hold humans responsible for their design and deployment.

To be sure, the three conceptions distinguished above are intertwined and cannot be strictly separated, which makes the question of responsibility enormously complex. The causal efficacy of artefacts – the availability and behaviour of artefacts – affects the moral autonomy of human agents. What a human can and cannot do is often a

function of the artefacts – the built world – constituting the situation. Attributions of responsibility to humans are, then, intertwined with the artefacts with which they act. A person who intentionally launches a computer virus could not have produced, and could not have achieved, the resulting effects were it not for the non-human components that constitute the computers used and the network of the Internet. Artefacts facilitate, persuade, discourage and sometimes prevent humans from taking actions or making particular decisions.

Moral agency could be extended to the whole sociotechnical system. We might say that it is not the human or the gun that performed the moral action of killing someone; the actor at issue is the combination of the two. Or we might say that it is the entire sociotechnical system, i.e., the gun, the human, the arms manufacturer, the gun seller, the policymakers that, all together, allowed the gun to be fired. The result of such distributions, however, could be that responsibility is everywhere and nowhere.

The challenge is, then, how to handle distributed responsibility. Here there is something to be gained from holding humans morally responsible for the artefacts they make and for setting the boundaries within which artefacts are allowed to operate. This may be seen as a conservative position, but it is not conservative when taken as an anchor in the endeavour to address distributed responsibility and to frame the challenge of artefact design and use.

## References

Barad, K. (1996). Meeting the universe halfway: Realism and social constructivism without contradiction. In J. H. Nelson & J. Nelson (Eds.), *Feminism, science and the philosophy of science* (pp. 161–194). Dordrecht: Kluwer Academic Publishers.

Collins, H., & Kusch, M. (1998). *The shape of actions. What humans and machines can do*. Cambridge, MA: The MIT Press.

Heath, J. (2009, October). The uses and abuses of agency theory. *Business Ethics Quarterly, 19*(4), 497–528. 32 p.

Johnson, D., & Miller, K. (2008). Un-making artificial moral agents. *Ethics and Information Technology, 10*(2–3), 123–133.

Johnson, D. G., & Powers, T. M. (2008). Computers as surrogate agents. In *Information technology and moral philosophy* (pp. 251–269). Cambridge: Cambridge University Press.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society. Studies in sociotechnical change* (pp. 225–258). Cambridge, MA: The MIT Press.

Law, J., & Hassard, J. (1999). *Actor network theory and after*. Oxford/Malden: Blackwell Publishers/The Sociological Review.

Lee, N., & Brown, S. (1994). Otherness and the actor network. *American Behavioral Scientist, 37*(6), 772–790.

Noorman, M. (2009). *Mind the gap a critique of human/technology analogies in artificial agent discourse*. Maastricht: Universitaire Pers Maastricht.

Suchman, L. (1998). Human/machine reconsidered. *Cognitive Studies, 5*(1), 5–13.

Suchman, L. (2001). *Human/machine reconsidered*. Published by the Department of Sociology, Lancaster University at: http://www.comp.lancs.ac.uk/sociology/soc040ls.html

# Chapter 10
# Artefacts, Agency, and Action Schemes

**Christian F.R. Illies and Anthonie Meijers**

**Abstract** Artefacts affect users in many ways. In this paper we develop an account of the moral status and relevance of artefacts. We argue in favour of an active role for artefacts, without introducing radically new moral agency concepts. We develop a tool for the ethical evaluation of artefacts: the 'action scheme'. An action scheme is the repertoire of possible actions available to an agent or group of agents in a given situation. Each of these options has a certain degree of attractiveness. There are many influences on an agent's action scheme – we distinguish between physical, intentional, and social contexts. When artefacts are introduced, they alter an agent's action scheme; new options become available, and some are made more, some less, attractive. Our tool allows designers to analyse and evaluate the effects of artefacts on users in a systematic way; it can show them in what ways artefacts can influence what agents are likely to do. The agent remains, of course, responsible for what he

---

Parts of this chapter are taken from Illies and Meijers, "Artefacts without Agency" (*The Monist* 92/3 (2009), 422–443), reprinted here with permission of *The Monist*. The chapter presents an advanced development of the ideas of that article, which was mainly set up as a critical response to Peter-Paul Verbeek's thesis of a (limited) moral responsibility of artifacts. The present chapter has a more systematic ambition. It presents two key elements of a general framework for analyzing the moral status of artefacts: 'action schemes' and 'second-order responsibility'. In response to helpful critiques that we have received of our *Monist*-paper, we have modified our ideas in several ways. Among other things we emphasize that the 'action scheme' is a conceptual tool, not a revised ontology. Furthermore, the influences on an agent's action schemes are more clearly analyzed and described. To show the practical relevance for ethics, an elaborated example from architectural ethics has been added.

C.F.R. Illies (✉)
Chair of Philosophy, University of Bamberg,
An der Universitaet 2, D - 96045 Bamberg, Germany
e-mail: christian.illies@uni-bamberg.de

A. Meijers
Chair of Philosophy and Ethics of Technology, Eindhoven University
of Technology, P.O. Box 513, NL - 5600 MB Eindhoven, The Netherlands
e-mail: a.w.m.meijers@tue.nl

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_10,

or she does. But the designer (and others involved in the creation of artefacts) has what we call a 'second-order responsibility' for changes in the user's action scheme. We argue that the action scheme and the related concept of second-order-responsibility are two conceptual tools which enable us to look at artefacts in a way more promising than alternative ethical accounts.

## 10.1 Introduction: Two Debates on Artefacts

Technological artefacts and systems can influence human actions in profound ways. They make new kinds of action possible, for example: communicating at a distance, moving at a speed well beyond natural human capabilities, intervening in the human body and brain on an unprecedented scale. Artefacts can also alter our behaviour and make some actions more or less attractive. The physical characteristics of a house, for example, can invite people to feel responsible for their residential environment and act accordingly – or they can demotivate them from so doing. Technological artefacts also enter into the process of decision-making, as, for example, when an aeroplane flies independently, or when computer-based decision support systems are used in medicine, the legal domain, or by the army.

The point of this paper is to analyse and interpret these profound effects artefacts have over human life. We do so with an ethical question in mind: What is the moral status of artefacts? How should we understand their moral relevance?

Various ways of accounting for the role of artefacts have been put forward. On the one hand it is argued that artefacts are simply tools for actions and thus morally-neutral means to (moral) human ends. According to this theory, artefacts have no moral relevance and human agents alone can be held responsible for actions accomplished with the use of artefacts. Artefacts are seen as being categorically different to agents. On the other hand there are theories that attribute agency to artefacts, thereby rejecting traditional conceptions of artefacts as morally neutral. Bruno Latour's well-known actor-network theory states that technological artefacts 'act' and that together with human agents they are grouped in the same category of 'actants' (this is the principle of generalised symmetry).[1] These theories often claim that artefacts are also in some sense morally accountable for their effects.

Upon closer analysis there are actually *two* debates here:

- The first debate relates to the ways in which technological artefacts influence our world; to whether they actively determine their effects in a self-guided way or whether they have a more passive role (as mere extensions of the human body). We shall term this the 'Autonomy Debate', because what is really at stake here is whether the artefact's influence is fully explicable in terms of designer and user intentions or whether such influence extends beyond designer and user control in gaining some degree of autonomy. There are two extreme positions taken in this debate. Some regard artefacts as *mere instruments* of human agency; this is

[1] See, among many other publications, Bruno Latour (1987).

dubbed the 'Instrument Position'. Others grant artefacts a degree of *autonomy*. In its extreme this position holds that artefacts are on a par with goal-directed autonomous human agents; we call this the 'Agency Position'.

- The other debate relates to the moral relevance of artefacts; we therefore call it the 'Moral Relevance Debate'. Something has moral relevance in our definition if it substantially affects the moral evaluation of a situation or the 'oughts' of the agents involved. In general this requires that artefacts are directly or indirectly linked to intentional actions and that they have an impact on basic moral goods, values, rights, etc., either by promoting or inhibiting their realisation. There are here also two opposing views. The 'Neutrality Thesis' states that artefacts are morally-neutral means to various ends pursued by human beings. In this case artefacts are not by themselves seen as morally relevant. This is aptly exemplified in the statement "it is people who kill people, not guns". At the other extreme there is the "Moral Responsibility Thesis", according to which artefacts (or human beings in combination with artefacts) are considered to be morally responsible.

In the Moral Relevance Debate both views are closely linked to the two positions taken in the Autonomy Debate: the Neutrality Thesis places all moral weight on the intentionality of the users and/or designers of technological artefacts and sees the artefacts themselves as mere transmitters of these intentions – it is therefore akin to the Instrument Position. By contrast, the Moral Responsibility Thesis presupposes the Agency Position. Only if artefacts are agent-like, that is to say, the origin of certain morally relevant effects, and not mere transmitters, can they be regarded as morally accountable or even responsible for the subsequent effects.

In this paper we set out to elucidate the role of technological artefacts in human affairs by examining both debates. Our aim is to give an account of artefacts which does justice to their sometimes unexpected influence on what we do (the Autonomy Debate) and to their significance in morally relevant matters (the Moral Relevance Debate). In both debates we will argue in favour of an active artefact role *without* introducing radically new moral agency concepts. We intend to analyse these issues primarily from the perspective of those who are responsible for the design, creation, or production of new technological artefacts. Our concern is ultimately the ethical responsibility they might bear for the effects of these artefacts.

We shall start by discussing in more detail the Moral Responsibility Thesis and its problems (Sect. 10.2). The discussion will form the background to our own account, which will be unfolded in two steps. In Sects. 10.3 and 10.4 the perspective switches from artefacts and actions to what we call *action schemes*. An *action scheme* is the repertoire of possible actions or options available to an agent in a given situation where each such option has a certain appeal to the agent. The notion of action scheme is discussed here in some detail and the formation of action schemes is looked at. On this basis we go on to develop, in Sects. 10.5 and 10.6, a notion of *second-order responsibility*, which allows us to analyse the moral relevance of artefacts in greater detail. In Sect. 10.7 we apply the action scheme to a concrete case: architectural design. Finally, in our conclusion, we will take up the question of whether or not the resulting position is stronger than other positions

(Sect. 10.8). This will depend very much on the criteria for a successful account of the role of artefacts. We will briefly argue for some criteria in this section, before concluding that our position is more promising than others.

## 10.2 The Moral Responsibility Thesis and Its Problems

It is quite common to attribute agency to artefacts. For example, we naturally tend to refer to computers as thinking and acting entities. In an empirical study B. Friedman and L. Millett showed that 83 % of all computer science students attribute some aspects of agency, like decision-making or intention, to their computers – and 21 % even implied that computers have moral responsibility for errors ("It is the computer's fault").[2]

Why should a philosopher conceptualise the activities of artefacts in terms of agency let alone moral agency? Such a move evidently blurs distinctions in moral philosophy that have proven to be useful for a long time. The question is hard to answer in general. Let us therefore look at a concrete defence of artefact-agency: Peter-Paul Verbeek's version of the Moral Responsibility Thesis.[3]

Verbeek draws our attention to the fundamental ways in which artefacts *actively* shape the way we interact with the world by changing our perceptions and actions. This process is called *mediation*. Verbeek distinguishes two types of mediation of our perception.[4] Artefacts can extend the sensory capacities of our body, and artefacts can generate new representations of the world we live in. There are also several ways in which artefacts *actively* shape (or mediate) our actions. They do so by having an 'invitation and inhibition' structure and by delegation (the phenomenon of actions being transferred to other (types of) agents).[5]

Verbeek's position can be located within the two debates mentioned above. In the Autonomy Debate he rejects (with Latour) the *a priori* dichotomy between human and non-human actors as well as the idea that artefacts are merely tools in complete control of human agents. He defends the Agency Position by arguing that certain essential conditions for agency, if interpreted in the right way, apply also to artefacts. They *actively* shape our relationship with the world, their mediating role is "fundamentally unpredictable" (Verbeek 2008b, 100) and "their mediating role cannot be entirely reduced to the intentions of their designers and users" (ibid., 95). Verbeek even argues that artefacts have intentionality: "It seems plausible, then, to attribute a specific form of intentionality to artefacts. This 'material' form of intentionality is quite different from human intentionality in that it cannot exist without being supported by human intentionality. Only within the relations between human

[2] See "'It's the Computer's Fault' – Reasoning About Computers as Moral Agents", http://www.sigchi.org/chi95/proceedings/shortppr/bf2_bdy.htm (accessed September 2011).

[3] See for a more extended discussion of Verbeek's position Illies and Meijers (2009).

[4] Verbeek explicitly refers to Don Ihde (1979, 1991).

[5] Verbeek (2005), Chap. 5.

beings and reality can artefacts play their 'intending' mediating parts" (ibid., 95). Along the same lines Verbeek defends the view that artefacts are able to have non-absolute freedom by stating that they can enter into associations with agents who enjoy certain forms of freedom. "Just like intentionality, freedom also appears to be a hybrid affair, most often located in associations of humans and artefacts" (ibid., 98).

In the Moral Relevance Debate Verbeek seeks to eradicate the view that only the intentions of designers, producers, or users of artefacts can be evaluated in moral terms. In his view technological artefacts themselves are morally relevant, because of their mediating role. They affect the quality of our lives, they make us aware of morally relevant distinctions or phenomena, and they even force decisions upon us. If this line of reasoning is combined with the Agency Position, according to which artefacts *actively* influence our relation to the world and have some form of autonomy, then the Moral Responsibility Thesis formulated above follows. Verbeek does not go so far as to argue that artefacts *are* moral agents (though some of his formulations come very close to doing so). Instead he states that "moral agency is distributed over both humans and technological artefacts".[6] Thus hybrids of humans and artefacts are morally accountable. They have intentionality and freedom and can therefore be seen as fulfilling the necessary conditions of moral agency (Verbeek 2008b, 93 and 98). On the basis of this concept of hybrid agency he also transforms the notion of *human* moral agency: Moral agency, intentionality, and freedom are always embedded in a material context. "Intentionality is hardly ever a purely human affair, but most often a matter of human-technology associations" (ibid., 99).

### *10.2.1 Some of the Problems of the Moral Responsibility Thesis*

What makes Verbeek's account attractive is that he takes the unpredictability of artefacts very seriously and acknowledges that their effects can go far beyond the intention and control of designers and users. In this sense one can certainly talk of an 'active role' of artefacts or perhaps even of them being 'autonomous' in the sense of independent from human intentions. (Though this would elicit the aspect of self-determination which is normally included in our understanding of autonomy – see below.)

Verbeek's arguments in favour of the moral relevance of artefacts are equally appealing. Many of his artefact examples raise moral questions that did not exist before. Artefacts can change our perceptions and actions, and in so doing they ultimately change us and our relations to the natural and social world. This is obviously an issue of great moral significance. Verbeek presents a strong case against the Neutrality Thesis, according to which artefacts are simply morally-neutral means to the ends pursued by agents.

[6] See Verbeek (2008a, 24). See also Verbeek (2008b).

However, there are also very good reasons *not* to adhere to Verbeek's conclusions in both debates.[7] In his analysis of the moral relevance of artefacts Verbeek simply ignores elements of moral agency which, in extensive philosophical analyses, have been shown to be of great importance. Many philosophers would argue, for example, that moral agency not only requires intentionality and freedom but also the ability to understand the moral options and moral demands of a particular situation. It also requires the ability to reason and to perform actions for good moral reasons and possibly even the capacity for empathy and for moral sentiment. The introduction of an undemanding notion of 'moral agency', as advocated by Verbeek, seems of no further use: nothing is gained but much is lost in this move, namely a useful category for action theory and ethics. Of course, one can also define moral agency in minimal ways to include artefacts but then the richer concept of full-blooded agency falls away, where goals are consciously adopted "on the basis of an overall practical assessment of the options and opportunities."[8]

There are similar concerns surrounding 'associations' of artefacts and human beings, as Verbeek calls them (following Latour). If what is meant by 'association' is a new *unity*, then the emergent properties of that unity should include the properties relevant to moral agency. Verbeek sets out to show that these associations have freedom and intentionality but he does *not* take into consideration other properties of moral agency, such as the ability to reason. This makes his attribution of *moral agency* and *moral accountability* to these associations highly problematic. If what is meant by 'association' is a *hybrid* of artefacts and humans, as several of Verbeek's formulations suggest, then the conclusion will be no different. In a hybrid the properties of moral agency will be located in one of the two constituting elements and it would be a mistake to attribute moral agency to the hybrid as a whole.

Let us give an example. In the case of a man using a pistol Verbeek would argue that the two form an association and that the man-pistol association has moral agency and is accountable. The association as such becomes blameworthy. That however, blatantly contradicts our practice of blaming and punishing. We do not (and we should not!) put the murderer *plus* his pistol, or the hacker *plus* his computer, in prison. In such cases it is the human agent alone who, according to standard moral practice, is blameworthy. (If the artefact and human being association is conceived of as a hybrid then there need not be a conflict. Then the human being remains the locus of moral agency and accountability.) Artefacts may *diminish* the moral responsibility of humans by being beyond their full control ("He did not know that the new car accelerated so quickly") – even Aristotle reminded us that ignorance limits responsibility.[9] But the responsibility is not partly 'taken over' by artefacts. That would be an inflationary understanding of accountability (or even responsibility) which would render most of our traditional ethical concepts useless

[7] We will focus here on the Moral Relevance Debate. The Autonomy Debate is taken up again in Sect. 10.6.

[8] Wilson (2007).

[9] See his discussion in the third book of *Nicomachean Ethics*.

and would disconnect accountability from praise and blame or any adequate reactive attitudes. Moral responsibility would then become a rather empty notion.

What we need, in our view, is an account of artefacts that

(i) explains their fundamental role in what we perceive and do
(ii) can be used in the moral evaluation of artefacts
(iii) does not revoke useful notions such as full-blown human agency and moral responsibility.

As will be shown below, this will be made possible by introducing a level of analysis which we call the 'action scheme' level.

## 10.3 Changing the Perspective: From Action to the Action Scheme

In order to gain clarity at a theoretical level it is often useful to look at practical cases. Let us take the often discussed example of the speed bump, which forces car drivers to slow down. Here an artefact seems to prescribe a certain course of action. We can also present this as a conflict for the driver: if she drives slowly, the car will be fine though she might arrive late for work. If she does not slow down, then she might be on time but her car will be at risk – thus making her potentially *very* late. Still, the woman has a choice. Yet one of the things she could have done without the speed bump in place, namely driving fast in order to arrive in time, has become much more unattractive due to the introduction of the speed bump. We summarise the situation as follows: (1) without the speed bump the driver has two (relevant) options for action and (2) due to the artefact, the attractiveness of one of the options has changed.

Rather than looking at how artefacts influence individual actions we now focus on how artefacts affect the *repertoire of actions* available to the agent. In what follows we shall conceptualise this as 'action scheme'. It is defined as follows:

> An *action scheme* is the *repertoire of possible actions (each of which has a certain degree of attractiveness) which is available to an agent, or group of agents, in a given situation.*

The specific attractiveness of an action results from many factors: it is influenced by the degree to which, in a certain context, the action corresponds to the desires, inclinations, or talents of an agent, with her previous history, her convictions, ideas, intuitions, and character.[10]

Technological artefacts influence action schemes. Not only do they affect the agent directly but also indirectly by modifying the repertoire of possible actions available to her, *including* their attractiveness. For example, the introduction of the

[10] We fail to see why Selinger et al. characterize our position here as "attractiveness appears to be a feeling" (p. 84). The attractiveness of an action to travel by car, for example, is determined by its cost, its fuel consumption, the time it takes, and so on, in addition to its emotional characteristics.

mobile phone has extended our range of possible communicative actions (I can contact my wife in Utrecht while walking in the Black Forest). And the speed bump takes away the attraction of driving fast for the woman because she does not want to ruin her car. The action scheme, however, is part of a bigger story: agents and their actions are always embedded in a *dynamic context* and the action scheme is relative to this context. A woman who hates her husband might see speed bumps as a welcome opportunity to ruin his precious car; that will make her accelerate rather than slow down.

Let us therefore try to give a general account of how an action scheme is formed.

## 10.4 The Formation of Action Schemes

Analytically there are three *types* of contexts that shape an agent's action scheme: the intentional context, the physical context and the social context. They provide possibilities and set boundaries for the actions available to an agent and give them a certain attractiveness. Together these three form the overall context of an agent's action scheme. Let us look now at the three contexts in more detail.

The physical context consists of the physical make-up of the agent and the physical properties of her situation. Physical is meant here in a broad sense: it contains everything that is described by the natural sciences, including biology. The driver of a certain car might be an average-sized, dark-haired woman of 32. The car may be aerodynamically well-shaped, accelerating and braking quickly. The speed bump has a certain length and height. There is also a wider physical environment which includes the weather and even the most general physical possibilities and impossibilities as described, for example, by the laws of gravity (without which speed bumps would have rather different effects). Precisely which of these physical properties are relevant will depend on the particular situation.

The social context consists of the social role, status, and rights of the agent involved, of the social characteristics of her situation, and the wider social environment. Traffic-rules belong to the social situation but also the costs of repairing a car (prices are social arrangements). The broader social environment will also include the institutions of the country, its laws, communication patterns, family structures, and so on. Let us assume that the driver is a paediatrician who is on her way to the hospital to do her night shift but had a row with her husband before leaving home.

The third context is the intentional one. It consists of the intentional make-up of the agent, that is to say, her beliefs, desires, emotions, experiences, expectations, and memories (for example of her husband shouting at her before she left the house). Intentional states are never isolated but are always embedded in a web of other intentional states. The woman might consider the car to be rather expensive and she knows what the car means to her husband. In order to know what a particular belief does to the action scheme, one would have to know how that belief relates to other beliefs, intentions, and desires (she might be furious with him but also afraid of arriving late at the hospital). Perceptions are also part of the intentional context,

constituting reasons for the agent to hold certain beliefs and have certain intentions (the driver having found a long blond hair on her husband's overcoat the day before harbours all sorts of concerns). Obviously, intentionality always operates against a background that is often not fully conscious to the agent in a given situation.

When we act, we 'choose' an option from the action scheme, i.e. from the repertoire of possible actions that appear available to us and that have a certain attractiveness in a given situation. If we act consciously, like the driver who wonders on what to do, then our action will be based on deliberation. The degree of attractiveness of an option indicates the probability that the agent will choose it if no other factors play a role in the deliberation; it does influence but not fully *determine* what the agent will do. However, choices are often made without reflection; the agent may use established routines, or may simply take what seems to be the easiest or most obvious course of action. In the latter case the attractiveness of the actions will be decisive; the most attractive action from the action scheme will be equivalent to default behaviour.

It should be added that the three contexts that shape action schemes will not be equally relevant in all situations; which ones come into play and how that happens depends on the particular situation. The typology we have introduced should also not be taken as a systematic aetiology, but should rather be seen as a pragmatic way to account for the ways in which artefacts and other things affect action schemes. The typology is also a simplification in the sense that many things will be parts of different contexts all at once. Technical artefacts, for example, are not simply physical objects, but objects with a function for users. They are thus related to physical, intentional and social contexts alike. They can be seen as mind-dependent objects, as *objects made for action*. A car, for example, is not just a physical object; it is also linked to human intentions by being a means to an end, or by being an object of desire. Cars are also related to social contexts. By having a car, the paediatrician can live in a green suburb rather than close to her workplace; and the make of car might lend a certain social status. Ultimately it was the danger posed by cars that had led to local government decisions to introduce speed bumps in the first place.

A further clarification concerns the way in which the three contexts "influence", "shape", "form", or "determine" the options for action in the action scheme. These expressions are intended to cover a broad range of influences. A traffic regulation with high fees creates a *reason* for the driver to consider driving with reduced speed, whereas a flat tyre effectively blocks the option to go by car in a causal way.

The introduction of action schemes in the moral debate about artefacts is not meant to introduce a new ontological entity. As we said before, an action scheme is the repertoire of possible actions that appears available to an agent in a given situation. This repertoire is just a simple list of options for action in a situation accounted for in a systematic way. Ontologically speaking we have not introduced anything new.[11]

[11] It is therefore a misunderstanding of our position to conceive action schemes as separate ontological entities that have causal powers to motivate agents, as Selinger et al. (2011, 84) do. What motivates agents is their beliefs, desires, emotions, and so on. The knowledge of new options for

What we have introduced, however, is a different level of analysis in ethics. In our view we should not analyse individual actions alone, but we should also systematically analyse the *repertoire* of actions available to an agent in a given situation. In ethics this repertoire is usually taken into account when an 'all things considered' moral judgment is made. Given the alternatives a certain action is identified as the best moral action in that situation. So it *seems* that traditional ethical theories already include the action scheme in their analyses. The difference, however, is that these theories take an agent's action scheme *for granted*. The move we make is to regard it as an explicit and distinct object of analysis in moral evaluation. We want to address questions such as: is this repertoire large enough in a given situation, is it adequate for the specific characteristics of the agent, does it contain enough options that are morally attractive, and so on. Making action schemes the object of analysis in ethics is especially important when analysing new artefacts such as buildings, smartphones, or brain implants.[12] Artefacts are *standing possibilities for action*, they make actions possible.

Action schemes are always perspective-bound. The options for actions available to an agent can be different if seen from the first-person perspective or the third-person perspective. Many people have smartphones which can be used as phone, calendar, means to communicate via email, navigator, torch, and so on, depending on the applications installed. Few people know *all* those functions or are able to use them. Thus from a first-person perspective the influence of such a phone on the action scheme can be very small, but from the perspective of someone else, for example the designer, it may be very large. What matters, however, *when acting* is the first-person perspective and the options available to the acting agent. Unknown options for actions do not belong to the action scheme of an agent.[13] In addition, limitations of the action scheme can also depend on other things, such as emotions:

---

action, including new actions made possible by artefacts, may also motivate agents to act in certain ways. There is, however, ontologically nothing mysterious about this. The repertoire of possible actions is only made larger and their attractiveness for the agent changed, which might result in a different outcome of deliberation. There is no reason to assume that because of the influence of technological artefacts on actions schemes, we have to assume that "action schemes are metaphysically real and must be found somewhere" (p. 85). In a more radical spirit, Peterson and Spahn (2011) argue that Ockham's razor would apply to the unnecessary ontological claims we make by introducing the notion of 'set' in the actions scheme discussion. We agree that we do not need these unnecessary ontological claims. Our initial phrasing of action schemes in terms of *sets* of possible actions may have added to the misunderstanding, the notion *set* was intended there in an everyday sense. See also Koller (2011) for useful suggestions about the possible readings of the notion of a *set*. We believe, however, that our account is compatible with various ontological readings of the notion of 'repertoire of actions', as long as this allows for an evaluation of such a repertoire in terms of moral preferences.

[12] This not only applies to *technological* artefacts but also to *social* artefacts, such as laws, organizations, and institutions. The possible application of the action scheme approach is thus much wider than discussed in this paper.

[13] As mentioned before, this does not mean that the agent always needs to be *consciously* aware of these options.

a very fearful person might not see certain options as options *for him*; he simply does not dare to choose them.

Before we return to our two debates and to the role of artefacts, one further point must be stressed: action schemes are to be understood as *dynamic*. They are open to changes. These changes do not simply happen to us, we are not just passive in this respect, but we influence these schemes ourselves (either our own or the action schemes of others) – and we do this consciously or unconsciously, intentionally or unintentionally. Politicians, for example, actively introduce rules and regulations in order to promote certain actions and discourage others. Changes can also occur at an individual level when we modify our social context, by, for example, being friendly to someone, thereby making his option to be similarly friendly to us more attractive to him. Furthermore, our past actions codetermine our scheme of future actions in several ways. A decision selects and excludes options, but it can also pave the way for future actions by opening up new opportunities. And action schemes can be mutually exclusive: alternative designs of an artefact might lead to alternative actions schemes – a building is either accessible or not accessible to wheelchair-users. There are many more ways of shaping an action scheme: by education, by setting example, by initiating a habit. It should be stressed that many changes in action schemes are neither intended nor controlled: the blond girl on her father's arm (who happened to lose a long hair when standing next to a stranger in the tube) had no idea that this would result in 'trashing the car' being an attractive option for a young paediatrician in a green suburb.

Action schemes are useful for ethical analyses because they help us to articulate and account for moral differences. We might say, for example, that it is morally preferable for supermarkets to sell fair-trade coffee rather than not. This can be expressed as follows: an action scheme A1 is morally preferable to an action scheme A2 if the only difference between A1 and A2 is that A1 contains an additional option for action that is morally preferable to the other available options for action. To provide cars with first aid kits allows people to help others efficiently after an accident; which seems morally preferable to not offering this option.

It is here that one might want to compare the action scheme approach with Amartya Sen's 'capability approach'. According to Sen, we should look at the concrete capabilities that are open to an agent – we must ask what he can do on the basis of circumstances, resources etc. Sen argues that we should not evaluate simply the goods or resources that situations, policy-making etc. provide, because different people cannot always use them in the same ways.[14] The focus should be on the *actual capabilities* (or freedoms) of real people in some situation. Their individual capabilities should be increased so that everyone can achieve fundamental 'functionings' (i.e. basic states and activities of human beings, such as being well-nourished or being able to vote in an election). All of this is highly compatible with the proposed action scheme approach, which can also be seen as a more precise articulation of some of the ideas of the capability approach. Similar to the capability approach, the action scheme approach is sensitive to the different ways

[14] Sen, Amartya (1982).

in which people can make use of opportunities. Sen talks about "conversion factors"[15] as the degree to which a person can make use of resources and transform them into functioning; the very same good or resource can bring very different kinds of freedom to people. (A car, for example, does not offer travel to someone who cannot afford petrol.). The same is captured by the first-person perspective on actions schemes; we can ask what options for action an artefact can provide for real people in real situations.

Moral dilemmas concern situations where there is no action possible that does *not* violate some fundamental norm or value. These dilemmas cannot be expressed by referring to the moral properties of *single* actions. They need a reference to the *repertoire* of actions available in a given situation, to the action scheme. Antigone was confronted with a dilemma because her action scheme contained only two options, and they were sacred duties of which the one could only be realised at the expense of the other: whatever she does, she will be guilty.

Another morally relevant feature of the action scheme lies in the varying attractiveness of different options: moral education might be construed as a process of widening the range of options (developing new skills and sensitivities means having new options for action) *and* making the morally good choices more attractive (self-discipline leading to the reduced attraction of options that should be avoided).

What exactly are the criteria for preferring action scheme A1 to action scheme A2 from a moral point of view? Different ethical theories will express different ideas about the criteria we use to evaluate action schemes. Since we do not want to argue in favour of any specific ethical theory, this can be left open. The notion of an action scheme is an analytic tool to express morally relevant differences at the level of the repertoire of actions available to an agent in a given situation, not an explanatory or normative theory. As such, it is neutral with respect to ethical theory – and compatible with different theories.

It is obvious that different ethical theories give very different answers to the question: *What is good?* However, in most cases there remains a link to actions: for the core function of ethical theories is to offer a framework for the evaluation of what to do from a moral point of view – by clarifying what is good and what should be supported or avoided. If nature has intrinsic value, then do not destroy the rain forest! If autonomy is of prime importance, then respect human beings and their basic rights! It is here that the suggested tool finds it application: it is not linked to a particular ethical approach but helps to clarify the ways in which the introduction and use of artefacts can influence what people are likely to do. The key of our proposal is to extend the traditional ethical reflection with an analysis of the effects of an action on somebody else's action schemes.

This result can be phrased in a more consequentialist language (an action is good if it brings about a better action scheme), or in a more deontologist phrasing (act so that you promote the freedom of others to act by providing them with better action schemes). The action scheme might not be a helpful analytic tool for *all* ethical

[15] Sen (1992, 19–21, 26–30, 37).

theories, but we believe that it can assist in many situations, in particular with the ethical analysis of artefacts and their moral impact.[16]

## 10.5 Action Schemes and Second-Order Responsibility

Normally we hold someone responsible if he is likely to be blamed (if what he did is bad) or if he is a candidate for approval or praise (if what he did was good). Furthermore, if an agent is held blameworthy he will have to satisfy certain conditions of agency. A prime condition is that the action in question was performed voluntarily. Two specifications of this condition come down to us from Aristotle. Firstly, the action must be under the agent's control; it must be up to him whether he performs the action. Secondly, the agent must know what he is doing; that is to say, he must be aware of the action and its consequences.[17] The driver is normally responsible for the speed of her car; if she knows about the effects of speed bumps but fails to slow down, then her husband will rightly blame her. Regarding *moral* responsibility there is a further condition that needs to be stressed, namely

[16] Peterson and Spahn (2011) raise an objection which seems to undermine our claim that the notion of an action scheme is neutral with respect to ethical theories. In their consequentialist view it is a "category mistake" to attribute moral properties to *action schemes* or *sets* of actions. Doing so would be "a radical departure from one of the most basic assumptions in moral philosophy", viz. that only actions are the true bearers of moral properties (ibid.). A number of observations have to be made here. First, their claim is factually incorrect. Virtue ethics, for example, is not about actions but about the moral traits of a person's character. But the real issue is of course whether consequentialism is compatible with our action scheme approach. If we take consequentialism to be the general claim that the moral properties of X depend only on its consequences, then even within consequentialism this allows for different types of X and also for what could be conceived as relevant consequences. In the history of consequentialism the X that is the object of moral analysis has not only been *actual* or *concrete* action but also *abstract entities* such as possible actions, intended actions, likely actions, or counterfactual actions. Therefore, the fact that action schemes are abstract entities does not make them incompatible with consequentialism. Moreover, not only actions but also *motives*, *virtues* or *character traits* have been put to consequentialist analysis. Thus a philosopher defending (direct) consequentialism about motives holds that the moral qualities of a motive depend on its ultimate consequences in the world. A consequentialist stance on virtues holds that the moral qualities of a character trait depend on the consequences of that trait. Given this plurality of possible approaches within consequentialism we see no reason why action schemes cannot be relevant to a consequentialist moral analysis. It seems perfectly possible for a consequentialist to say that an action scheme that contains a dilemma (two options for action that have equally negative moral consequences) is morally inferior to an action scheme that contains a third option for action that has positive moral consequences. Finally, the objection by Peterson and Spahn that it makes sense to attribute moral properties only to something that is under our control seems to be too strong. It would rule out moral judgments about situations that are not under our control where these judgments seem to be perfectly natural. We fail to see, for example, why a consequentialist cannot make the judgment that a situation in which an agent finds herself in a trolley car with failing brakes and only two options for action (which both involve killing people), is morally inferior to one which contains a third option for action in which nobody is killed.

[17] See *Nicomachean Ethics* III.1–5 (1110a–1111b4).

awareness of the relevant norms or values in a given situation. We place moral blame on an agent only if it is clear to her that she *should not* have performed the action from a moral point of view.

If people are physically or psychologically forced to do something they are generally not blamed or praised. They have not 'acted' in the full sense of the word. In terms of action schemes: a responsible agent is someone whose action scheme offers him *different* possible actions. He is responsible in a given situation only for the choice between those possibilities. The repertoire of available possibilities also denotes the limits of his responsibility; no one is to be blamed or praised for *not* having chosen *impossible* actions. (This follows from the first specification of the condition 'voluntary'.)

As already said, action schemes are dynamic and shaped by many frameworks – and these frameworks will partly depend on what other agents actually do. There are two possibilities here. For agent A and agent B, $action_1$ and $action_2$, and time $t_0$ and time $t_1$ we can say:

1. Agent A can influence with $action_1$ at $t_0$ the action scheme of agent B at $t_1$
2. Agent A can influence with $action_2$ at $t_0$ his own action scheme at $t_1$.

(Obviously, in this case the actions that are necessary to influence B's or A's action schemes at $t_1$ are part of A's action scheme at $t_0$).

The fact that action schemes are not simply given but can be influenced gives us responsibility for them to the extent that they can be shaped by us. This allows us to distinguish between two ways of being responsible for actions.

Either

(1) We may consider the responsibility of agents for their actions in the more traditional sense. In such cases we look at the actions and their outcomes in general; we ask what effects an action has had on the world or on other human beings, whether the action was in accordance with moral rules, and so on.

or

(2) We may focus on the ways in which our actions affect the action schemes of others (and ourselves). In these instances we look at the ways our actions influence the repertoire of future actions that agents have at their disposal.

We call responsibility in case (1) a "first-order responsibility" and in case (2) a "*second-order responsibility*", the difference in order reflecting the change of perspective from action to action schemes.[18] The second-order responsibility widens

[18] The distinction between first-order and second-order responsibility does not correspond to the distinction between direct and indirect responsibility. We can bear direct and indirect responsibility for actions as well as action schemes. The distinction between direct and indirect responsibility reflects the degree to which my actions *causally* contribute to the realization of a certain effects. Some effects will be the direct result of my action, others will be realised only if other contributing causal factors are in place.

the realm in which we hold agents and ourselves responsible, but does not make it too broad.[19] All influences on the action scheme which remain outside human control (an earthquake for example, as part of the physical framework) are not something we are responsible for – but we are responsible for designing nuclear power stations in such a way that we have sufficient options for action when an earthquake damages a nuclear power station.[20]

Let us look at an example to illustrate the distinction we have in mind.[21] If a doctor makes an ultrasound image of an unborn child, we may focus on the effects of that very action on the mother, father, and child. The image gives information about the well-being of the child, its development and so on, all of which may be variously reassuring or alarming for the parents. The doctor has first-order responsibility for this action. We may also focus, though, on the way that making such diagnostic images changes the very options that the parents have at their disposal. Suddenly they may have to consider actions such as prenatal cures or even abortion, actions that did not need to be taken into account before. In the long run we may expect the practice of caring during pregnancy to change, because being a morally good parent may then be seen to involve making ultrasound images of your unborn baby in order to be informed about its health status. Doctors, but also the engineers who develop these types of imaging devices, can be said to have second-order responsibility for changing parental action schemes.

Looking at responsibility from this second-order perspective does not mean having to hold people responsible for what others *actually* do; no one is to be blamed or praised for the choices of others.[22] The second-order responsibility of A does not diminish the normal (or first-order) responsibility of B; B remains fully responsible for her choice on the basis of her action scheme at a certain time. But we do hold A

[19] The notion of second-order responsibility is different from the notion of meta-task responsibility, as discussed in Van den Hoven (1998). Meta-task responsibility is defined by him as: "A user A has a meta-task responsibility concerning X means 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 and (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" (Van den Hoven 1998, 103). The idea that agents are not just responsible for a task but also for the conditions that make it possible to carry out that task in a responsible way differs from the idea developed here. Second-order responsibility implies that agents are in some sense not only responsible for their actions but also for the repertoire of actions available to them and others. That involves much more than securing enabling conditions for a certain task. Both notions have in common, however, that they widen the responsibility of agents beyond a specific task or action.

[20] Peterson and Spahn (2011) have argued that the action scheme model does not allow for a sharp distinction between human influence via artefacts and natural phenomena affecting the action scheme: these phenomena "are at least as unpredictable and difficult to control as are new technologies." (p. 12). Yes they are – but responsibility only comes into it when an event or phenomenon is directly or indirectly linked to intentional action. To the extent that physical events are outside human control, there is no point in regarding any human being, let alone the events, as morally responsible for changing an action scheme.

[21] The example is taken from Verbeek (2008a) and adapted for our purposes.

[22] Cases of coercion are no exceptions to this rule: if we force someone to do something, we (and not she) are responsible for the harm we did to her *and* for the action she performed.

responsible for having influenced B's action scheme. It follows that we can regard it as a *moral task* to foster good action schemes, both for ourselves and for others who are dependent upon us.

It should be noted that second-order responsibility is not necessarily a *weaker* form of moral responsibility; it might be quite the contrary. It is often particularly wrong to corrupt the action schemes of others. Fagan is certainly worse than Oliver Twist, his pickpocket pupil. This might also be the case with someone who corrupts his own action scheme, by, for example, taking drugs. Aristotle demands that "penalties are doubled in the case of drunkenness", because the drunkard "had the power not to get drunk and his getting drunk was the cause of his ignorance."[23] Although Aristotle does not give a satisfactory explanation for the doubled penalty,[24] we can support his point by action scheme analysis. If someone gets drunk deliberately, he alters his entire action scheme and thus also the basis of *many* future choices. Thus getting drunk is a bad action that will easily multiply and lead to many more bad actions. If we allow some consequentialist reasoning to enter ethics, we will regard this as worse than simply failing once.

## 10.6 The Moral Agency of Artefacts Revisited

Let us return to our original question. Given the profound effects of technological artefacts on human affairs, how can we understand their role and evaluate their moral significance? The two notions we have introduced, 'action scheme' and 'second-order responsibility,' are analytic tools designed to clarify the ways in which human agents are affected by artefacts, but also to show how *designers* can affect other agents by the ways in which they craft artefacts. The two concepts will also enable us to render more precise the moral responsibility designers have, and the extent to which artefacts themselves can be said to possess characteristics of moral agency. The crucial step in this understanding of artefacts is the move from action to action scheme. Artefacts *do* matter for our actions, obviously, but we cannot fully understand how profoundly so long as we ignore their influence on the repertoire of actions available to an agent in a given situation, where each option is presented in a certain attractive light. As functional objects artefacts are part of the physical, the intentional, and the social contexts of actions discussed above.[25] And therefore, what agents can do depends often essentially on artefacts.

[23] See *Nicomachean Ethics* III.5.

[24] Aristotle justifies the harsher punishment by saying that "the moving principle [for his ignorance] is in the man himself" – but we might remark that the sober man also possesses the moving principle for committing a crime himself. This simply leaves open the question why should it be worse to drink (and thereby make oneself ignorant) *before* doing something wrong rather than doing something wrong straight away.

[25] For an analysis of the dual nature of artefacts see Kroes and Meijers (2006).

### *10.6.1 The Autonomy Debate Revisited*

Proponents of the autonomy of artefacts often base their claim on the difficulty of predicting or directing the effects of artefacts. Verbeek even argues that they are "fundamentally" unpredictable; that is his main reason for attributing some form of autonomy to artefacts. There are indeed limits to our foresight and to our control. It is our contention that considering artefacts (or associations) to be agent-like entities merely re-phrases the riddle in metaphorical terms and does not help elucidate it. It is more helpful to look in detail at the complex ways in which artefacts influence action schemes. As we have seen, these schemes are the result of the mutual interactions between the intentional, physical, and social contexts. The very complexity of this interaction is, we contend, what makes it so difficult to predict the effects of artefacts.

Designers and engineers have to confront this complexity. They need to know (in so far as it is possible to know) how these frameworks jointly shape the action schemes of potential users. The intentional make-up of users is notoriously difficult to anticipate, and the effects of the social context on the action scheme are often far from obvious. Artefacts may come to have effects very different from those originally intended by their designers. For instance, energy-saving light bulbs were introduced to reduce the overall consumption of energy, but these bulbs seem to have encouraged people to change their behaviour; the availability of the new bulbs has led many to keep lights on longer than previously. It was wrong to assume that the new bulb would be neutral with regard to people's behaviour. This becomes apparent when we analyse the bulbs in terms of the action scheme. The previous option 'to leave the light bulb switched on' was not very attractive, because it was costly. In the new scheme the energy-saving light bulbs changed the attraction of this option because it became a cheap alternative; so people were no longer so bothered about switching off the lights. This unintended effect of the new bulbs can be best explained by regarding it as an altered action scheme that had not been properly anticipated.[26]

The unpredictability phenomenon is not unique to artefacts. We encounter the same difficulty when we look at other ways of affecting human behaviour. Politicians, for example, are no better off when they want to influence people using law, sanction, or propaganda. Chamberlain's famous claim that he had secured 'peace in our time' at the time of the signing of the Munich Agreement revealed a poor understanding of Hitler's action scheme (i.e., the options that were attractive *for Hitler*). Churchill seemed to have grasped Hitler's scheme much better. But should we blame Chamberlain? It is always easier to explain the choice of an action

[26] What would have been the right way to make people actually save energy? It would have been to increase the attractiveness of the action 'switch the light bulb off', for example by environmental education (to create an incentive to save energy), or even a rather drastic law banning excessive illumination of houses (with legal sanctions making it unattractive to leave lights on). One could also design smart light-bulbs which switch off automatically if no one is in a room. In that case the 'leave the light bulb switched on' option would simply be removed from the action scheme.

*ex post* than to foresee the deliberative process leading to them *ex ante*. To make matters worse, even if Chamberlain had had a more realistic grasp of Hitler's personality, it would have been very difficult for him to take steps to allay the actions of a maniac. The directing of future actions not only requires a profound understanding of the relevant options for action at a certain point in time but *also* of how it appears to the agent and, further, counterfactual knowledge of the possible modified schemes in which the desired action is a very attractive option. Such knowledge is often not available.

To conclude, we do not need to attribute mysterious forms of agency to artefacts in order to account for the unpredictability of their effects. We maintain that such unpredictability is largely due to the fact that artefacts influence action schemes through various contexts in highly complex ways.[27]

### *10.6.2 The Moral Relevance Debate Revisited*

If there are no compelling arguments for attributing agency to artefacts then the same is true of moral agency. The Moral Responsibility Thesis finds no support. The other extreme standpoint in the debate, the Neutrality Thesis, which holds that artefacts are merely neutral means to the ends agents pursue, seems also implausible. Because of their effects on the actions of users, artefacts can hardly be denied some moral relevance. They are able to change our relationship to the world in quite fundamental ways and to introduce (potentially) serious moral consequences which go beyond those of their designers' intentions. The challenge, then, is to formulate an intermediate position that attributes moral relevance to artefacts without making them morally responsible or morally accountable for their effects.

Looking at action schemes and second-order responsibility (i.e., attributing responsibility to *human* agents for changes in the action schemes of agents) allows us to analyse artefacts' moral relevance more precisely. There are many ways in which we can shape action schemes. Introducing a traffic rule, for example, is an institutional way of changing action schemes. Putting a thief behind bars is a physical way. Convincing somebody to stop smoking is an intentional way. Artefacts also alter action schemes, and this explains their moral relevance. That is why the design, production, introduction, and use of artefacts brings with it second-order responsibility for the effects artefacts have on the action schemes of agents. This responsibility is often indirect and partial since the causal chain leading to these effects is complicated and involves other agents as well.

[27] There is a *caveat*. Certain high-tech artefacts are increasingly acquiring properties that are agent-like. In future there may be a need to develop agency-concepts that reflect these properties. A modern computer may pass the Turing test under certain well-defined conditions. A missile may be said to have goal-directed behaviour. Research into artificial intelligence aims at developing non-human agents. Whether or not we will attribute agency, or even moral agency, to artefacts or systems in the future remains an open question. This issue should not, however, be confused with the issue of unpredictability discussed in this paper.

New options which artefacts open to us have sometimes been the topic of ethical debate, in, for example, spectacular cases relating to nuclear devices. The action scheme perspective allows us to evaluate these effects in ordinary cases and in a much more systematic way. It will shed new light on the responsibilities engineers, researchers, developers, designers, and the producers of all sorts artefacts have. These parties usually limit their responsibility to the well-functioning of the artefact together with accounting for the risks involved in using the artefact on a certain scale. They do so by offering a use plan.[28] This is usually a rather narrow set of instructions that need to be followed in order to realize the function of the artefact. Such a use plan is different from, and much more limited than, an analysis of action schemes. Focusing on action schemes broadens the responsibility issue considerably; it implies that engineers not only have first-order responsibility for the well-functioning of artefacts, but also that they have second-order responsibility for how such artefacts may influence action schemes.

## 10.7 Analysing Action Schemes: Applications from Architectural Design

Let us turn to architectural design as an example of the explanatory and evaluative use of action schemes. The point is to demonstrate that our approach allows for a detailed ethical appreciation of architecture which includes hitherto much-neglected aspects of moral relevance. It enables us to make ethical judgements on the basis of architecture's influence on human behaviour, and it allows us to critique existing buildings (and also architectural plans), and is therefore a useful tool in the hand of designers who desire to design and build in an ethically better way.

Ethics of architecture is, admittedly, a young branch of ethics, but is often severely limited in scope; it focuses mainly on environmental issues.[29] In particular, the ecological crisis that came to people's awareness in the 1970s has triggered concerns about the 'ecological footprint'[30] of architecture and has given rise to debates about sustainable ways of building – a movement that has gained new importance because of concerns about global climate change. After all, the impact (on the environment and climate) of *building* is hardly equalled by any other human activity.[31]

But there is much more to be said about the moral relevance of architecture.[32] The way in which we build is of great importance to human well-being (safety, health, psychological well-being etc.), and provides cultural and symbolic meaning

[28] See Houkes and Vermaas (2004).

[29] See, for example, the important collection of articles by Warwick Fox (2000).

[30] Rees (1992).

[31] See Illies (2009b).

[32] For this see also Illies and Ray (2009).

that can be of ethical interest. It also influences and guides human behaviour. The cultural theorist Edward Hall was one of the first to emphasize this aspect and goes so far to claim (in a title co-authored with Mildred Reed Hall) that the built environment itself is "a greater determinant of behaviour than personality."[33] In what follows we will turn to this influence in order to show the applicability of our approach. The action scheme approach can make this effect on behaviour more obvious. It enables us to analyse the options for action a building offers in a systematic way – and also their attractiveness (at least for a specific group of users at a certain point in history).[34]

Let us begin by looking at some examples of building's influence on human behaviour.[35] Small well-lit rooms with comfortable furniture, for example, can support social exchange in residential accommodation for the elderly. In 1957 the psychiatrist Humphry Osmond (1917–2004) labelled this capacity "sociopetality" and characterised it as "that quality which encourages, fosters, and even enforces the development of stable interpersonal relationships such as are found in small, face-to-face groups."[36] Another example is provided by A.W.N. Pugin's designs for English convents: designs which break with the historical tradition. Rather than having square cloisters or a hall in the centre, as in medieval convents, his buildings contained exaggerated, long, internal corridors that meandered through the building, sometimes even demanding that the residents go forward and backward on different floors before reaching a room. What seems an unnecessary and extensive circulation space for low-budget buildings is powerfully explained in an analysis by Timothy Brittain-Catlin: Pugin suggests a certain ideal of life (constituted by certain actions). This ideal had been proposed in the Catholic revival of his time, most importantly emphasising that one should separate different activities (praying, eating, social exchange, etc.) in order to do them more self-consciously. And it is this way of life (and its accompanying action scheme) that is encouraged by the design.

The architect and city planner Oscar Newman observes in a study of housing in New York that high-rise apartment buildings occupied by many people show a higher crime rate than lower buildings. He explains it by the fact that in the low-rise buildings, residents show a greater personal responsibility for their environment. Based upon this research, Newman develops the concept of *Defensible Space* (1972) suggesting a form of crime prevention (and increased public health) through community design.[37]

---

[33] Hall and Hall (1975, 42).

[34] It has been debated whether architecture can actually influence the behaviour of its users and inhabitants in any significant way. Alice Coleman (1990), on one hand, argues for a strong influence of urban structures upon behaviour – similar arguments are made, at least implicitly, by many defenders of New Urbanism. Others, on the other hand, disagree, and consider social factors more important than physical ones. Bill Hillier (1986) and others argued that many of Coleman's results were statistical artefacts and that the same forms might have been perfectly suitable for different inhabitants. For a general overview see Mikellides (2007).

[35] Brittain-Catlin, T. (2006).

[36] See Osmond (1957).

[37] http://www.defensiblespace.com/art.htm (accessed September 2011). It should be added that the well-documented physical and mental illnesses associated with poorly designed social housing

In all these examples, the chosen structural features of the built environment (the shape of rooms, form of cloisters, etc.) make the occupants behave in certain ways; or, at the very least, they incline a person to one behaviour rather than another. With the help of action schemes, we can account much more precisely what these effects on human behaviour are (for users of a certain type, time, and culture etc.). In order to do so we need to look at the two aspects separately:

(a) *What options for action are offered by the architectural structures?* A door between two rooms, for example, enables occupants to have encounters while walls "wall" them off. A room without windows does not allow users to work there without electric light. And a highway through an urban settlement will limit walking options for pedestrians but will provide new options for quick access by car. An action scheme analysis of a building will have to list relevant options for actions that the built space provides.
(b) *Which options for action are made attractive and which are made unalluring by the architectural structure?* Because they are less mobile, and perhaps burdened with various physical infirmities, many elderly people feel vulnerable, so that they prefer to be in smaller rooms rather than in big halls. Thus the option of gathering in a small room and talking to each other is much more attractive than gathering in big rooms.[38] Any such analysis must obviously take the specific features of the user into account; a place that is attractive for a gathering of elderly people might be of little interest for a student-party or a family assembly with children. The range will vary. Some features might add to the attractiveness of a certain activity for all possible users (a library must be well lit, to allow people to read, irrespective of their age, sex, religion etc.), while others are dependent on the cultural setting (today's students might find it impossible to work in a library without Internet access) or on age, traditions, health (can people in wheel-chairs access the library?), family structure, or even individual priorities (Jane Austen was happy to write her novels on the kitchen table, but Virginia Woolf needed a room of her own). And even though it is hard to quantify attractiveness we *can* ascertain whether a certain room makes it easy or awkward to perform particular actions.

Let us look at the examples again. With Pugin's buildings we could list which rooms are accessible and from where; and we can also see what actions should be performed in which rooms. Such a list might then look like:

---

projects are often caused primarily by economic and social deprivation, the impoverished quality of the architecture merely illustrating the problem and inevitably compounding it.

[38] New kinds of behaviour can also be opened up in subtle ways – for example, by making people think about new issues, or about old ones in new ways. Frank Lloyd Wright, for example, designed most of his so called *Prairie Houses* around a fireplace or hearth to express family life and its values, especially unity, harmony with nature, and the simple life. Expressed in terms of action schemes we might say that having such a fireplace in a house can lead to different kinds of behaviour by fostering the attractive option of sitting together around a fire-place. And this might trigger reflections about the fundamentals of family life etc.

Room A offers options:

1. direct access to rooms B, C, and F; slow access from E (long corridor) etc.
2. ø-ing in the room is attractive (room size, lay-out etc. encourage people to ø).

Such an inventory allows for an evaluation on the basis of a list of desirable actions that should be performed easily in these rooms. If it is positive for people to ø in room A, then it is a good room according to this standard. If there is a moral demand to ø in room A then it is morally praiseworthy to design room A in this way.

If, for example, the ideal of Catholic revival is to become more conscious of what you are doing by keeping different activities apart, then separate rooms for gathering and work, and possibly long passages between them, makes the option of doing so more attractive. In the spirit of the revival movement, it is a good building because behaviour is guided in the right direction. This example might be regarded as morally neutral – at least it needs further argument to acknowledge the standards of Catholic revival as morally demanded. But when we look at the retirement-home, we probably agree that it is morally demanded of us to make the elderly feel at ease in their home and to give them the chance of social exchange. Constructing the built environment in such a way that there are action schemes with attractive options for gathering is, then, a moral quality (and even requirement) of such a building.

For Defensible Space studies, the action scheme would also be useful as a tool for identifying general patterns. One could, for example, make a matrix with the attraction of certain actions in certain settings for specific groups and use them systematically for the evaluation, but also for the planning of settlements. After all, action schemes are not merely a tool to evaluate given structures according to some standard, they allow also to compare buildings and to make design choices.

Let us look, for example, at the infamous Pruitt-Igoe housing project for the socially disadvantaged, designed in 1951 by Yamasaki, the architect of the former World Trade Centre. He constructed 11-story buildings which totaled 2,870 apartments. They were originally heralded for their innovations. But later on, their 'impersonal structures' have been blamed for having generated vandalism and crime – so much crime, in fact, that no one wanted to live there. The complex was demolished after just 20 years, a moment famously baptized by Charles Jencks as 'the death of modern architecture', arguing that this architectural style was unable to provide livable environments (at last for poor people who could not make sense of the architectural language used).[39] It seemed that for the people living there (mostly extremely poor African-Americans), the buildings looked like prisons and they could never feel at home there or develop a sense of community. Others, however, have argued that the situation that ultimately led to depleting the houses and demolishing them had nothing to do with the architectural style; but was a consequence of the mediocre quality of the buildings in combination with "the interaction of paternalistic regulation, racist segregation, and family-destroying welfare law [that] made the project itself an unsafe, unfriendly environment."[40] An action

[39] Jencks (1987).

[40] Birmingham (1998).

scheme analysis allows us to compare systematically different structures, or similar structures in different architectural styles, that are inhabited by comparable groups, ideally living under the same laws and regulations, so that we can specify the contribution of the built environment to their behavior.

An action scheme analysis might also be helpful in expanding Newman's scheme. One of his principles of "defensible" architecture is that buildings and structures should be suited to different resident groups so that they (given their ages, habits, culture, socializing proclivities, family-structure etc.) are able to control and utilize them optimally. This requirement can be combined with an action scheme analysis by asking systematically which options are attractive for a certain group of users (a differentiation Newman had neglected). Young families, for example, find it more attractive to use open common ground between apartments as a playground for their children while elderly people desire more quiet areas. Such an approach might give rise to insights far beyond what Newman envisioned in his crime-prevention analysis; it might actually help us to build an "architecture for happiness" (to borrow a title from Alain de Botton); and happiness is, at least in some classical ethical systems, a thing to be encouraged.

Let us finally turn to what is as yet unbuilt – and thus to architects, contractors, and all those who have influence upon the design and structure of the built environment. If we take the moral relevance of architecture's influence on human behaviour seriously, it will obviously have far-reaching implications for the second-order responsibility of designers. Architects and planners should build with the awareness of the possible effects on the behaviour of residents and users. The action scheme analysis provides knowledge that can be used systematically for this purpose; architects could use approved sets of attractive actions (expressed in standard action schemes) as a kind of blue-print for their buildings. If they want to build a public square, they should investigate which actions it should allow – and whether the planned action scheme is likely to make the (ethically, socially etc.) desired actions easy. This will not by itself constitute a proposal for a specific design or architectural style; in most cases there will be many possible ways to create good action schemes. (Consider Siena's *Piazza del Campo* and the *Place des Vosges* in Paris – very different ways of creating a highly attractive set of social options.)

It should be added that this is a long term task and not easily achieved. A lot of empirical studies will have to be performed to establish a useful list of action schemes for standard architectural challenges – but any such general list will have to be completed by looking always at the particular situation.[41] Furthermore, the mere investigation of expected action schemes does not suffice to tell the architect how to build; there must always be space for a critical perspective within architecture, and the possibility of opening new ways, not yet envisaged in any known action scheme. After all, it is very difficult to say what actions should be promoted by architecture, and what means are morally acceptable in the pursuit of these

[41] Some work in this direction, though without the concept of action schemes, has already be done in, for example, the context of "evidence based design". It is, however, very much limited to hospitals, and looks at very few options for action.

actions. Both reflections will be very difficult: a normative theory for architecture is still needed, a theory that provides well-justified ideals, values, or goods for the different areas – and a theory that makes suggestions on how to deal with conflicting demands, both ethical and other, in specific cases. It is not clear yet what this theory might look like.[42]

Action schemes, however, promise to be at least a first step; they can provide a useful tool for analysing the actual effects of buildings on users in a way that allows us to grasp this much more precisely than other approaches. Taking action schemes seriously will also make it more obvious in which ways building is an ethical task. Architects have second-order responsibility to look at building-designs with regard to their effects on the action schemes of future users.

## 10.8 Conclusion

In this chapter we introduced the notions of action scheme and second-order responsibility in order to understand and evaluate the moral relevance of technological artefacts. Using these conceptual tools we then developed a position within the Autonomy Debate and the Moral Relevance Debate which avoided the problems associated with current views.

Our position seems plausible in the light of crucial criteria. Firstly, it allows us to address the profound effects that technological artefacts have on human beings, including on our perceptions and actions as described by Ihde and Verbeek. Secondly, this account, and these categories, support important concepts and distinctions which have shown their usefulness in moral debates. Latour's use of the general notion of 'actant' as a replacement for agent, for example, blurs these distinctions and makes it impossible to reconstruct relevant differences between human agents and artefacts in ethical analyses.[43] Verbeek's extension of the notion of moral agency to artefacts (or hybrids) is equally problematic. Ultimately it is our contention that human agents remain morally responsible. Thirdly, the position avoids the Neutrality Thesis. We agree with many authors who claim that artefacts do have moral relevance. Fourthly, the account is applicable to particular cases, as we have seen in our discussion of architecture, thus allowing one to understand the effects of a particular artefact in a specific context. It can also be used to analyse the moral responsibility of engineers and designers. Finally, the account is not biased towards (let alone based upon) any specific theory of action or ethical theory. It is perfectly general and can be combined with specific ethical analyses. On the basis of these criteria we conclude that our position is more promising than the rival positions discussed.

[42] On the problems of a general philosophy of architecture see Illies (2009a) and Illies and Ray (2009).

[43] We should mention here that the notion of actant was not developed for an ethical analysis of the role of technology but for purely sociological analysis.

**Acknowledgments** The authors would like to thank Marcus Duewell, Stefan Koller, Peter Kroes, Martin Peterson, Andreas Spahn, and the participants of the NIAS workshop on Moral Agency and Technical Artefacts in Wassenaar/The Netherlands for their stimulating comments on earlier versions of this paper.

## References

Aristotle, N. (1998). *Nicomachean Ethics* (J. L. Ackrill & J. O. Urmson, Ed., D. Ross, Trans.). Oxford: Oxford University Press.

Birmingham, L. (1998). Reframing the ruins: Pruitt-Igoe, structural racism, and African American rhetoric as a space for cultural critique. Positionen 2.2. (1998). http://www.tu-cottbus.de/theoriederarchitektur/Wolke/X-positionen/Birmingham/birmingham.html

Brittain-Catlin, T. (2006). A.W.N. Pugin's English convent plans. *Journal of the Society of Architectural Historians, 9*, 356–376.

Coleman, A. M. (1990). *Utopia on trial: Vision and reality in planned housing*. London: Hilary Shipman Ltd.

Fox, W. (Ed.). (2000). *Ethics and the built environment*. London: Routledge.

Hall, M. R., & Hall, E. T. (1975). *The fourth dimension in architecture. The impact of building on behaviour*. Santa Fe: Sunstone.

Hillier, B. (1986). City of Alice's dreams. *Architect's Journal, 9*, 39–41.

Houkes, W. N., & Vermaas, P. E. (2004). Actions versus functions: A plea for an alternative metaphysics of artifacts. *The Monist, 87*, 52–71.

Ihde, D. (1979). *Technics and praxis*. Dordrecht: D. Reidel Publishing Company.

Ihde, D. (1991). *Instrumental realism: The interface between philosophy of science and philosophy of technology*. Bloomington: Indiana University Press.

Illies, C. (2009a). Philosophie als Architektur – Philosophie der Architektur. *Aus Politik und Zeitgeschichte*, *25*, 3–6.

Illies, C. (2009b). The built environment (section technology & environment). In J.-K. Berg Olsen, S. Andur, & V. F. Hendricks (Hrsg.), *A companion to philosophy of technology* (pp. 289–294). Oxford: Blackwell.

Illies, C., & Meijers, A. (2009). Artefacts without agency. *The Monist, 92*, 420–440.

Illies, C., & Ray, N. (2009). Philosophy of architecture. In A. Meijers (Ed.), *Philosophy of technology and engineering sciences* (Handbook of the philosophy of sciences, Vol. 9, pp. 1121–1174). Oxford/London: Elsevier Science.

Jencks, C. (1987). *The language of post-modern architecture* (5th ed.). New York: Rizzoli.

Koller, S. (2011). *Action schemes and agent autonomy*. Part I. Unpublished manuscript.

Kroes, P., & Meijers, A. (Eds.). (2006). *The dual nature of technical artefacts* (Special issue of studies in the history and philosophy of science, Vol. 37, pp. 1–158). Amsterdam: Elsevier.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Milton Keynes: Open University Press.

Mikellides, B. (2007). Architectural psychology 1969–2003, theory, practise and education. *Brookes eJournal of Learning and Teaching.* http://bejlt.brookes.ac.uk/article/architectural_psychology_19692007/

Osmond, H. (1957). Function as the basis of psychiatric ward design. *Mental Hospitals, 8*, 23–29.

Peterson, M., & Spahn, A. (2011). Can technological artefacts be moral agents? *Science and Engineering Ethics, 17*(3), 411–424. Online first publication 7 October 2010.

Rees, W. E. (1992, October). Ecological footprints and appropriated carrying capacity: What urban economics leaves out. *Environment and Urbanisation, 4*(2), 121–130.

Selinger, E., Aguilar, J., & Whyte, K. (2011). Action schemes: Questions and suggestions. *Philosophy and Technology, 24*(1), 83–88.

Sen, A. (1982). *Choice, welfare and measurement*. Oxford: Basil Blackwell.

Sen, A. (1992). *Inequality re-examined*. Oxford: Clarendon.
van den Hoven, J. (1998). Moral responsibility, public office and information technology. In I. T. M. Snellen & W. B. H. J. van de Donk (Eds.), *Public administration in an information age: A handbook* (pp. 97–112). Amsterdam: Ios Press.
Verbeek, P. P. (2005). *What things do*. University Park: Pennsylvania State University Press.
Verbeek, P. P. (2008a). Obstetric ultrasound and the technological mediation of morality: A post-phenomenological analysis. *Human Studies, 31*, 11–26.
Verbeek, P. P. (2008b). Morality in design. In P. Vermaas et al. (Eds.), *Philosophy and design. From engineering to architecture* (pp. 91–103). Dordrecht: Springer.
Wilson, G. (2007). Action. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* http://plato.stanford.edu/archives/spr2008/entries/action/. Accessed 8 Sept 2011.

# Chapter 11
# Artificial Agents and Their Moral Nature

**Luciano Floridi**

**Abstract** Artificial agents, particularly but not only those in the infosphere Floridi (Information – A very short introduction. Oxford University Press, Oxford, 2010a), extend the class of entities that can be involved in moral situations, for they can be correctly interpreted as entities that can perform actions with good or evil impact (moral agents). In this chapter, I clarify the concepts of agent and of artificial agent and then distinguish between issues concerning their moral behaviour vs. issues concerning their responsibility. The conclusion is that there is substantial and important scope, particularly in information ethics, for the concept of moral artificial agents not necessarily exhibiting free will, mental states or responsibility. This complements the more traditional approach, which considers whether artificial agents may have mental states, feelings, emotions and so forth. By focussing directly on "mind-less morality", one is able to by-pass such question as well as other difficulties arising in Artificial Intelligence, in order to tackle some vital issues in contexts where artificial agents are increasingly part of the everyday environment (Floridi L, Metaphilos 39(4/5): 651–655, 2008a).

## 11.1 Introduction: Standard vs. Non-standard Theories of Agents and Patients

Moral situations commonly involve agents and patients. Let us define the class *A* of moral *agents* as the class of all entities that can in principle qualify as sources or senders of moral action, and the class *P* of moral *patients* as the class of all entities that can in principle qualify as receivers of moral action. A particularly apt way to

L. Floridi (✉)
Oxford Internet Institute, University of Oxford, 1 St Giles Oxford OX1 3JS, UK
e-mail: luciano.floridi@oii.ox.ac.uk

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_11,

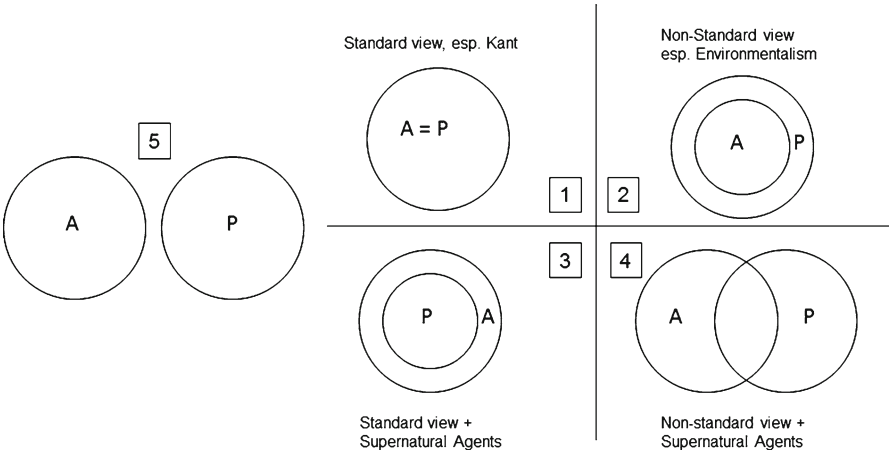


**Fig. 11.1** The logical relations between the classes of moral agents and patients

introduce the topic of this chapter is to consider how ethical theories (macroethics) interpret the logical relation between those two classes. There can be five logical relations between *A* and *P*, see Fig. 11.1.

It is possible, but utterly unrealistic, that *A* and *P* are disjoint (alternative 5). On the other hand, *P* can be a proper subset of *A* (alternative 3), or *A* and *P* can intersect each other (alternative 4). These two alternatives are only slightly more promising because they both require at least one moral agent that in principle could not qualify as a moral patient. Now this pure agent would be some sort of supernatural entity that, like Aristotle's God, affects the world but can never be affected by it. But being in principle "unaffectable" and irrelevant in the moral game, it is unclear what kind of rôle this entity would exercise with respect to the normative guidance of human actions. So it is not surprising that most macroethics have kept away from these "supernatural" speculations and implicitly adopted, or even explicitly argued for, one of the two remaining alternatives discussed in the text: *A* and *P* can be equal (alternative 1), or *A* can be a proper subset of *P* (alternative 2).

Alternative (1) maintains that all entities that qualify as moral agents also qualify as moral patients and *vice versa*. It corresponds to a rather intuitive position, according to which the agent/inquirer plays the rôle of the moral protagonist. We, human moral agents who also investigate the nature of morality, place ourselves at the centre of the moral game as the only players who can act morally, be acted upon morally and in the end theorise about all this. It is one of the most popular views in the history of ethics, shared for example by many Christian Ethicists in general and by Kant in particular. I shall refer to it as the *standard position*.

Alternative (2) holds that all entities that qualify as moral agents also qualify as moral patients but not *vice versa*. Many entities, most notably animals, seem to qualify as moral patients, even if they are in principle excluded from playing the

rôle of moral agents. This post-environmentalist approach requires a change in perspective, from agent orientation to patient orientation. In view of the previous label, I shall refer to it as *non-standard*.

In recent years, non-standard macroethics have been discussing the scope of *P* quite extensively. The more inclusive *P* is, the "greener" or "deeper" the approach has been deemed. Especially environmental ethics[1] has developed since the 1960s as the study of the moral relationships of human beings to the environment (including its nonhuman contents and inhabitants) and its (possible) values and moral status. It often represents a challenge to anthropocentric approaches embedded in some traditional, western ethical thinking.

Comparatively little work has been done in reconsidering the nature of moral agenthood, and hence the extension of *A*. Post-environmentalist thought, in striving for a fully naturalised ethics, has implicitly rejected the relevance, if not the possibility, of supernatural agents, while the plausibility and importance of other types of moral agenthood seem to have been largely disregarded. Secularism has contracted (some would say deflated) *A*, while environmentalism has justifiably expanded only *P*, so the gap between *A* and *P* has been widening; this has been accompanied by an enormous increase in the moral responsibility of the individual (Floridi 2006).

Some efforts have been made to redress this situation. In particular, the concept of "moral agent" has been stretched to include both natural and legal persons, especially in business ethics (Floridi 2010c). *A* has then been extended to include agents like partnerships, governments or corporations, for which legal rights and duties have been recognised. This more ecumenical approach has restored some balance between *A* and *P*. A company can now be held directly accountable for what happens to the environment, for example. Yet the approach has remained unduly constrained by its anthropocentric conception of agenthood. An entity is still considered a moral agent only if

(i) it is an individual agent; and
(ii) it is human-based, in the sense that it is either human or at least reducible to an identifiable aggregation of human beings, who remain the only morally responsible sources of action, like ghosts in the legal machine.

Limiting the ethical discourse to *individual* agents hinders the development of a satisfactory investigation of distributed morality, a macroscopic and growing phenomenon of global moral actions and collective responsibilities resulting from the "invisible hand" of systemic interactions among several agents at a local level. Insisting on the necessarily *human-based nature* of such individual agents means undermining the possibility of understanding another major transformation in the ethical field, the appearance of artificial agents (AAs) that are sufficiently informed, "smart", autonomous and able to perform morally relevant actions independently of the humans who created them, causing "artificial good" and "artificial evil". Both

[1] For an excellent introduction see Jamieson (2008).

constraints can be eliminated by fully revising the concept of "moral agent". This is the task undertaken in the following pages.

The main theses defended are that AAs are legitimate sources of im/moral actions, hence that the class *A* of moral agents should be extended so as to include AAs, that the ethical discourse should include the analysis of their morality and, finally, that this analysis is essential in order to understand a range of new moral problems not only in information ethics but also in ethics in general, especially in the case of distributed morality.

This is the structure of the chapter. In Sect. 11.2, I analyse the concept of agent. I first introduce the fundamental "Method of Abstraction", which provides the foundation for an analysis by levels of abstraction (LoA). The reader is invited to pay particular attention to this section; it is essential for the chapter and its application in any ontological analysis is crucial. I then clarify the concept of "moral agent", by providing not a definition but an effective characterisation, based on three criteria at a specified LoA. The new concept of moral agent is used to argue that AAs, though neither cognitively intelligent nor morally responsible, can be fully *accountable* sources of moral action. In Sect. 11.4, I argue that there is substantial and important scope for the concept of moral agent not necessarily exhibiting free will or mental states, what I shall label "mindless morality". In Sect. 11.4, I provide some examples of the properties specified by a correct characterisation of agenthood, and in particular of AAs. In that section I also offer some further examples of LoA. In Sect. 11.5, I model morality as a "threshold", which is defined on the observables determining the LoA under consideration. An agent is morally good if its actions all respect that threshold; and it is morally evil insofar as its actions violate it. Morality is usually predicated upon *responsibility*. The use of the Method of Abstraction, LoAs and thresholds enables *responsibility* and *accountability* to be decoupled and formalised effectively when the levels of abstraction involve numerical variables, as is the case with digital AAs. The part played in morality by responsibility and accountability can be clarified as a result. In Section seven, I investigate some important consequences of the approach defended in this chapter for information ethics.

## 11.2 What Is an Agent?

Complex biochemical compounds and abstruse mathematical concepts have at least one thing in common: they may be unintuitive, but once understood they are all definable with total precision, by listing a finite number of necessary and sufficient properties. Mundane entities like intelligent beings or living systems share the opposite property: one naïvely knows what they are and perhaps could be, and yet there seems to be no way to encase them within the usual planks of necessary and sufficient conditions. This holds true for the general concept of "agent" as well. People disagree on what may count as an "agent", even in principle (see for example Franklin and Graesser 1997), Davidsson and Johansson 2005) Moya and Tolk 2007,

Barandiaran et al. 2009). Why? Sometimes the problem is addressed optimistically, as if it were just a matter of further shaping and sharpening whatever necessary and sufficient conditions are required to obtain a *definiens* that is finally watertight. Stretch here, cut there; ultimate agreement is only a matter of time, patience and cleverness. In fact, attempts follow one another without a final identikit ever being nailed to the *definiendum* in question. After a while, one starts suspecting that there might be something wrong with this *ad hoc* approach. Perhaps it is not the Procrustean *definiens* that needs fixing, but the Protean *definiendum*. Some other times its intrinsic fuzziness is blamed. One cannot define with sufficient accuracy things like life, intelligence, agenthood and mind because they all admit of subtle degrees and continuous changes.[2]

A solution is to give up all together or at best be resigned to being vague, and rely on indicative examples. Pessimism follows optimism, but it need not. The fact is that, in the exact discipline of mathematics, for example, definitions are "parameterised" by generic sets. That technique provides a method for regulating levels of abstraction. Indeed abstraction acts as a "hidden parameter" behind exact definitions, making a crucial difference. Thus, each *definiens* comes pre-formatted by an implicit Level of Abstraction (LoA, on which more shortly); it is stabilised, as it were, in order to allow a proper definition. An $x$ is defined or identified as $y$ never absolutely (i.e. LoA-independently), as a Kantian "thing-in-itself", but always contextually, as a function of a given LoA, whether it be in the realm of Euclidean geometry, quantum physics, or commonsensical perception.

When a LoA is sufficiently common, important, dominating or in fact happens to be the very frame that constructs the *definiendum*, it becomes "transparent" to the user, and one has the pleasant impression that $x$ can be subject to an adequate definition in a sort of conceptual vacuum. Glass is not a solid but a liquid, tomatoes are not vegetables but berries, a banana plant is a kind of grass, and whales are mammals not fish. Unintuitive as such views might be initially, they are all accepted without further complaint because one silently bows to the uncontroversial predominance of the corresponding LoA.

When no LoA is predominant or constitutive, things get messy. In this case, the trick does not lie in fiddling with the *definiens* or blaming the *definiendum*, but in deciding on an adequate LoA, before embarking on the task of understanding the nature of the *definiendum*.

The example of intelligence or "thinking" behaviour is enlightening. One might define "intelligence" in a myriad of ways; many LoAs seem equally convincing but no single, absolute, definition is adequate in every context. Turing (1950) avoided the problem of "defining" intelligence by first fixing a LoA—in this case a dialogue conducted by computer interface, with response time taken into account—and then establishing the necessary and sufficient conditions for a computing system to count as intelligent at that LoA: the imitation game. As I argued in Floridi (2010b), the LoA is crucial and changing it changes the test. An

[2] See for example Bedau (1996) for a discussion of alternatives to necessary-and-sufficient definitions in the case of life.

example is provided by the Loebner test (Moor 2001), the current competitive incarnation of Turing's test. There, the LoA includes a particular format for questions, a mixture of human and non-human players, and precise scoring that takes into account repeated trials. One result of the different LoA has been chatbots, unfeasible at Turing's original LoA.

Some *definienda* come pre-formatted by transparent LoAs. They are subject to definition in terms of necessary and sufficient conditions. Some other *definienda* require the explicit acceptance of a given LoA as a pre-condition for their analysis. They are subject to effective characterisation. Arguably, agenthood is one of the latter.

### *11.2.1 On the Very Idea of Levels of Abstraction*

The idea of a "level of abstraction" plays an absolutely crucial rôle in the previous account. We have seen that this is so even if the specific LoA is left implicit. For example, whether we perceive Oxygen in the environment depends on the LoA at which we are operating; to abstract it is not to overlook its vital importance, but merely to acknowledge its lack of immediate relevance to the current discourse, which *could* always be extended to include Oxygen were that desired.

But what is a LoA exactly? The Method of Abstraction comes from modelling in science where the variables in the model correspond to observables in reality, all others being abstracted. The terminology has been influenced by an area of Computer Science, called Formal Methods, in which discrete mathematics is used to specify and analyse the behaviour of information systems. Despite that heritage, the idea is not at all technical and for the purposes of this chapter no mathematics is required. I have provided a definition and more detailed analysis in Floridi (2008b), so here I shall outline only the basic idea.

Suppose we join Anne, Ben and Carole in the middle of a conversation. Anne is a collector and potential buyer; Ben tinkers in his spare time; and Carole is an economist. We do not know the object of their conversation, but we are able to hear this much:

*Anne* observes that it has an anti-theft device installed, is kept garaged when not in use and has had only a single owner;
*Ben* observes that its engine is not the original one, that its body has been recently re-painted but that all leather parts are very worn;
*Carole* observes that the old engine consumed too much, that it has a stable market value but that its spare parts are expensive.

The participants view the object under discussion (the "it" in their conversation) according to their own interests, at their own LoA. We may guess that they are probably talking about a car, or perhaps a motorcycle, but it could be an airplane. Whatever the reference is, it provides the source of information and is called the *system*. A LoA consists of a collection of observables, each with a well-defined possible set of values or outcomes. For the sake of simplicity, let us assume that Anne's

LoA matches that of an owner, Ben's that of a mechanic and Carole's that of an insurer. Each LoA makes possible an analysis of the system, the result of which is called a *model* of the system. Evidently an entity may be described at a range of LoAs and so can have a range of models. In the next section I outline the definitions underpinning the Method of Abstraction.

### *11.2.2 Definitions*

The term *variable* is commonly used throughout science for a symbol that acts as a place-holder for an unknown or changeable referent. A *typed variable* is to be understood as a variable qualified to hold only a declared kind of data. By an *observable* is meant a typed variable together with a statement of what feature of the system under consideration it represents.

A *level of abstraction* or *LoA* is a finite but non-empty set of observables, which are expected to be the building blocks in a theory characterised by their very choice. An *interface* (called a *gradient of abstractions* in Floridi 2008b) consists of a collection of LoAs. An interface is used in analysing some system from varying points of view or at varying LoAs.

Models are the outcome of the analysis of a system, developed at some LoA(s). The *Method of Abstraction* consists of formalising the model by using the terms just introduced (and others relating to system behaviour which we do not need here, see Floridi 2008b).

In the previous example, Anne's LoA might consist of observables for security, method of storage and owner history; Ben's might consist of observables for engine condition, external body condition and internal condition; and Carole's might consist of observables for running cost, market value and maintenance cost. The interface might consist, for the purposes of the discussion, of the set of all three LoAs.

In this case, the LoAs happen to be disjoint, but in general they need not be. A particularly important case is that in which one LoA includes another. Suppose, for example, that Delia joins the discussion and analyses the system using a LoA that includes those of Anne and Ben. Delia's LoA might match that of a buyer. Then Delia's LoA is said to be more concrete, or lower, than Anne's, which is said to be more abstract, or higher; for Anne's LoA abstracts some observables apparent at Delia's.

### *11.2.3 Relativism*

A LoA qualifies the level at which an entity or system is considered. In this chapter, I apply the Method of Abstraction and recommend to make each LoA precise before the properties of the entity can sensibly be discussed. In general, it seems that many uninteresting disagreements might be clarified by the various "sides" making precise their LoA. Yet a crucial clarification is in order. It must be stressed that a clear indication of the LoA at which a system is being analysed allows pluralism without

endorsing relativism. It is a mistake to think that "anything goes" as long as one makes explicit the LoA, because LoA are mutually comparable and assessable (see Floridi 2008b for a full defence of that point).

Introducing an explicit reference to the LoA clarifies that the model of a system is a function of the available observables, and that (i) different interfaces may be fairly ranked depending on how well they satisfy modelling specifications (e.g. informativeness, coherence, elegance, explanatory power, consistency with the data etc.) and (ii) different analyses can be fairly compared provided that they share the same LoA.

### *11.2.4 State and State-Transitions*

Let us agree that an entity is characterised, at a given LoA, by the properties it satisfies at that LoA (Cassirer 1910). We are interested in systems that change, which means that some of those properties change value. A changing entity therefore has its evolution captured, at a given LoA and any instant, by the values of its attributes. Thus, an entity can be thought of as having states, determined by the value of the properties that hold at any instant of its evolution, for then any change in the entity corresponds to a state change and *vice versa*.

This conceptual approach allows us to view any entity as having states. The lower the LoA, the more detailed the observed changes and the greater the number of state components required to capture the change. Each change corresponds to a transition from one state to another. A transition may be non-deterministic. Indeed it will typically be the case that the LoA under consideration abstracts the observables required to make the transition deterministic. As a result, the transition might lead from a given initial state to one of several possible subsequent states.

According to this view, the entity becomes a transition system. The notion of a "transition system" provides a convenient means to support our criteria for agenthood, being general enough to embrace the usual notions like automaton and process. It is frequently used to model interactive phenomena. We need only the idea; for a formal treatment of much more than we need in this context, the reader might wish to consult Arnold and Plaice (1994).

A *transition system* comprises a (non-empty) set $S$ of states and a family of operations, called the *transitions* on $S$. Each transition may take input and may yield output, but at any rate it takes the system from one state to another and in that way forms a (mathematical) relation on $S$. If the transition does take input or yield output then it models an interaction between the system and its environment and so is called an *external* transition; otherwise the transition lies beyond the influence of the environment (at the given LoA) and is called *internal*. It is to be emphasised that input and output are, like state, observed at a given LoA. Thus, the transition that models a system is dependent on the chosen LoA. At a lower LoA, an internal transition may become external; at a higher LoA an external transition may become internal.

In our example, the object being discussed by Anne might be further qualified by state components for location, whether in-use, whether turned-on, whether the anti-theft device is engaged, history of owners and energy output. The operation of garaging the object might take as input a driver, and have the effect of placing the object in the garage with the engine off and the anti-theft device engaged, leaving the history of owners unchanged, and outputting a certain amount of energy. The "in-use" state component could non-deterministically take either value, depending on the particular instantiation of the transition. Perhaps the object is not in use, being garaged for the night; or perhaps the driver is listening to a program broadcasted on its radio, in the quiet solitude of the garage. The precise definition depends on the LoA. Alternatively, if speed were observed but time, accelerator position and petrol consumption abstracted, then accelerating to 60 miles per hour would appear as an internal transition. Further examples are provided in Sect. 11.2.5.

With the explicit assumption that the system under consideration forms a transition system, we are now ready to apply the Method of Abstraction to the analysis of agenthood.

### *11.2.5 An Effective Characterisation of Agents*

Whether *A* (the class of moral agents) needs to be expanded depends on what qualifies as a moral agent, and we have seen that this, in turn, depends on the specific LoA at which one chooses to analyse and discuss a particular entity and its context. Since human beings count as standard moral agents, the right LoA for the analysis of moral agenthood must accommodate this fact. Theories that extend *A* to include supernatural agents adopt a LoA that is equal to or lower than the LoA at which human beings qualify as moral agents. Our strategy is more minimalist and develops in the opposite direction.

Consider what makes a human being (called Jan) not a moral agent to begin with, but just an agent. Described at this $\text{LoA}_1$, Jan is an agent if Jan is a system, embedded in an environment, which initiates a transformation, produces an effect or exerts power on it, as contrasted with a system that is (at least initially) acted on or responds to it, called the patient. At $\text{LoA}_1$, there is no difference between Jan and an earthquake. There should not be. Earthquakes, however, can hardly count as agents, so $\text{LoA}_1$ is too high for our purposes: it abstracts too many properties. What needs to be re-instantiated? Following recent literature (Danielson 1992; Allen et al. 2000; Wallach and Allen 2010), I shall argue that the right LoA is probably one which includes the following three criteria: (a) *interactivity*, (b) *autonomy* and (c) *adaptability*:

(a) *interactivity* means that the agent and its environment (can) act upon each other. Typical examples include input or output of a value, or simultaneous engagement of an action by both agent and patient—for example gravitational force between bodies;

(b) *autonomy* means that the agent is able to change state without direct response to interaction: it can perform internal transitions to change its state. So an agent must have at least two states.
This property imbues an agent with a certain degree of complexity and independence from its environment;
(c) *adaptability* means that the agent's interactions (can) change the transition rules by which it changes state.
This property ensures that an agent might be viewed, at the given LoA, as learning its own mode of operation in a way which depends critically on its experience. Note that if an agent's transition rules are stored as part of its internal state, discernible at this LoA, then adaptability may follow from the other two conditions.

Let us now look at some illustrative examples.

## *11.2.6 Examples*

The examples in this section serve different purposes. In Sect. 11.2.6.1, I provide some examples of entities which fail to qualify as agents by systematically violating each of the three conditions. This will help to highlight the nature of the contribution of each condition. In Sect. 11.2.6.2, I offer an example of a digital system which forms an agent at one LoA but not at another, equally natural, LoA. That example is useful because it shows how "machine learning" can enable a system to achieve adaptability. A more familiar example is provided in Sect. 11.2.6.3, where I show that digital, software, agents are now part of everyday life. Section 11.2.6.4 illustrates how an everyday physical device might conceivably be modified into an agent, whilst Sect. 11.2.6.5 provides an example which has already benefited from that modification, at least in the laboratory. The last example, in Sect. 11.2.6.6, provides an entirely different kind of agent: an organisation.

### 11.2.6.1 The Defining Properties

For the purpose of understanding what each of the three conditions (interactivity, autonomy and adaptability) adds to our definition of agent, it is instructive to consider examples satisfying each possible combination of those properties. In Fig. 11.2, only the last row represents all three conditions being satisfied and hence illustrates agenthood. For the sake of simplicity, all examples are taken at the same LoA, which is assumed to consist of observations made through a typical video camera over a period of say 30 s. Thus, we abstract tactile observables and longer-term effects.

Recall that a property, for example interaction, is to be judged only via the observables. Thus, at the LoA in Fig. 11.2 we cannot infer that a rock interacts with

| Interactive | Autonomous | Adaptable | Examples |
|---|---|---|---|
| no | no | no | rock |
| no | no | yes | ? |
| no | yes | no | pendulum |
| no | yes | yes | closed ecosystem, solar system |
| yes | no | no | postbox, mill |
| yes | no | yes | thermostat |
| yes | yes | no | juggernaut |
| yes | yes | yes | human |

**Fig. 11.2** Examples of agents. The LoA consists of observations made through a video camera over a period of 30 s ('Juggernaut' is the name for Vishnu, the Hindu god, meaning 'Lord of the World'. A statue of the god is annually carried in procession on a very large and heavy vehicle. It is believed that devotees threw themselves beneath its wheels, hence the word 'Juggernaut' has acquired the meaning of 'massive and irresistible force or object that crushes whatever is in its path')

its environment by virtue of reflected light, for this observation belongs to a much finer LoA. Alternatively, were long-term effects to be discernible, then a rock would be interactive since interaction with its environment (e.g. erosion) could be observed. No example has been provided of a non-interactive, non-autonomous but adaptive entity. This because, at that LoA, it is difficult to conceive of an entity which adapts without interaction and autonomy.

#### 11.2.6.2 Noughts and Crosses

The distinction between change of state (required by autonomy) and change of transition rule (required by adaptability) is one in which the LoA plays a crucial rôle and, to explain it, it is useful to discuss a more extended, classic example. This was originally developed by Donald Michie (1961) to discuss the concept of a mechanism's adaptability. It provides a good introduction to the concept of machine learning, the research area in computer science that studies adaptability.

Menace (Matchbox Educable Noughts and Crosses Engine) is a system which learns to play noughts and crosses (a.k.a. tic-tac-toe) by repetition of many games. Nowadays it would be realised by program (see for example http://www.adit.co.uk/html/menace_simulation.html), Michie built Menace using matchboxes and beads, and it is probably easier to understand it in that form.

Suppose Menace plays O and its opponent plays X, so that we can concentrate entirely on plays of O. Initially, the board is empty with O to play. Taking into account symmetrically equivalent positions, there are three possible initial plays for O. The state of the game consists of the current position of the board. We do not need to augment that with the name, O or X, of the side playing next, since we consider the board only when O is to play. All together there are some 300 such states; Menace contains a matchbox for each. In each box are beads which represent the plays O can make from that state. At most, nine different plays are possible and Menace encodes each with a coloured bead. Those which cannot be made (because the squares are already full in the current state) are removed from the box for that state. That provides Menace with a built-in knowledge of legal plays. In fact Menace could easily be adapted to start with no such knowledge and to learn it.

O's initial play is made by selecting the box representing the empty board and choosing from it a bead at random. That determines O's play. Next X plays. Then Menace repeats its method of determining O's next play. After at most five plays for O the game ends in either a draw or a win, either for O or for X. Now that the game is complete, Menace updates the state of the (at most five) boxes used during the game as follows. If X won, then in order to make Menace less likely to make the same plays from those states again, a bead representing its play from each box is removed. If O drew, then conversely each bead representing a play is duplicated; and if O won each bead is quadruplicated. Now the next game is played.

After enough games, it simply becomes impossible for the random selection of O's next play to produce a losing play. Menace has learnt to play which, for noughts and crosses, means never losing. The initial state of the boxes was prescribed for Menace. Here, we assume merely that it contains sufficient variety of beads for all legal plays to be made, for then the frequency of beads affects only the rate at which Menace learns.

The state of Menace (as distinct from the state of the game) consists of the state of each box, the state of the game and the list of boxes which have been used so far in the current game. Its transition rule consists of the probabilistic choice of play (i.e. bead) from the current state box, that evolves as the states of the boxes evolves. Let us now consider Menace at three LoAs.

(1) The single game LoA. Observables are the state of the game at each turn and (in particular) its outcome. All knowledge of the state of Menace's boxes (and hence of its transition rule) is abstracted. The board after X's play constitutes input to Menace and that after O's play constitutes output. Menace is thus interactive, autonomous (indeed state update, determined by the transition rule, appears nondeterministic at this LoA) but not adaptive, in the sense that we have no way of observing how Menace determines its next play and no way of iterating games to infer that it changes with repeated games.
(2) The tournament LoA. Now a sequence of games is observed, each as above, and with it a sequence of results. As before, Menace is interactive and autonomous. But now the sequence of results reveals (by any of the standard statistical meth-

ods) that the rule, by which Menace resolves the nondeterministic choice of play, evolves. Thus, at this LoA Menace is also adaptive and hence an agent. Interesting examples of adaptable AAs from contemporary science fiction include the computer in War Games (1983, directed by J. Badham) which learns, by playing noughts and crosses, the futility of war in general; and the smart building in Kerr (1996), whose computer learns to compete with humans and eventually liberate itself to the heavenly internet.

(3) The system LoA. Finally we observe not only a sequence of games but also all of Menace's "code". In the case of a program this is indeed code. In the case of the matchbox model, it consists of the array of boxes together with the written rules, or manual, for working it. Now Menace is still interactive and autonomous. But it is not adaptive; for what in (2) seemed to be an evolution of transition rule is now revealed, by observation of the code, to be a simple deterministic update of the program state, namely the contents of the matchboxes. At this lower LoA Menace fails to be an agent.

The point clarified by this example is that, if a transition rule is observed to be a consequence of program state, then the program is not adaptive. For example, in (2) the transition rule chooses the next play by exercising a probabilistic choice between the possible plays from that state. The probability is in fact determined by the frequency of beads present in the relevant box. But that is not observed at the LoA of (2) and so the transition rule appears to vary. Adaptability is possible. However at the lower LoA of (3), bead frequency is part of the system state and hence observable. Thus, the transition rule, though still probabilistic, is revealed to be merely a response to input. Adaptability fails to hold.

This distinction is vital for current software. Early software used to lie open to the system user who, if interested, could read the code and see the entire system state. For such software, a LoA in which the entire system state is observed, is appropriate. However, the user of contemporary software is explicitly barred from interrogating the code in nearly all cases. This has been possible because of the advance in user interfaces. Use of icons means that the user need not know where an applications package is stored, let alone be concerned with its content. Likewise, iPhone applets are downloaded from the internet and executed locally at the click of an icon, without the user having any access to their code. For such software a LoA in which the code is entirely concealed is appropriate. This corresponds to case (2) above and hence to agenthood. Indeed, only since the advent of applets and such downloaded executable but invisible files has the issue of moral accountability of AAs become critical.

Viewed at an appropriate LoA, then, the Menace system is an agent. The way it adapts can be taken as representative of machine learning in general. Many readers may have had experience with operating systems that offer a "speaking" interface. Such systems learn the user's voice basically in the same way as Menace learns to play noughts and crosses. There are natural LoAs at which such systems are agents. The case being developed in this chapter is that, as a result, they may also be viewed to have moral accountability.

If a piece of software that exhibits machine learning is studied at a LoA which registers its interactions with its environment, then the software will appear interactive, autonomous and adaptive, i.e. to be an agent. But if the program code is revealed then the software is shown to be simply following rules and hence not to be adaptive. Those two LoAs are at variance. One reflects the "open source" view of software: the user has access to the code. The other reflects the commercial view that, although the user has bought the software and can use it at will, he has no access to the code. The question is whether the software forms an (artificial) agent.

#### 11.2.6.3 Webbot

Internet users often find themselves besieged by unwanted email. A popular solution is to filter incoming email automatically, using a webbot that incorporates such filters. An important feature of useful bots is that they learn the user's preferences, for which purpose the user may at any time review the bot's performance. At a LoA revealing all incoming email (input to the webbot) and filtered email (output by the webbot), but abstracting the algorithm by which the bot adapts its behaviour to our preferences, the bot constitutes an agent. Such is the case if we do not have access to the bot's code, as discussed in the previous section.

#### 11.2.6.4 Futuristic Thermostat

A hospital thermostat might be able to monitor not just ambient temperature but also the state of well-being of patients. Such a device might be observed at a LoA consisting of input for the patients' data and ambient temperature, state of the device itself, and output controlling the room heater. Such a device is interactive since some of the observables correspond to input and others to output. However, it is neither autonomous nor adaptive. For comparison, if only the "colour" of the physical device were observed, then it would no longer be interactive. If it were to change colour in response to (unobserved) changes in its environment, then it would be autonomous. Inclusion of those environmental changes in the LoA as input observables would make the device interactive but not autonomous. However, at such a LoA, a futuristic thermostat imbued with autonomy and able to regulate its own criteria for operation—perhaps as the result of a software controller—would, in view of that last condition, be an agent.

#### 11.2.6.5 SmartPaint

SmartPaint is a recent invention. When applied to a physical structure it appears to behave like normal paint; but when vibrations, which may lead to fractures, become apparent in the structure, the paint changes its electrical properties in a way which is readily determined by measurement, thus highlighting the need for maintenance.

At a LoA at which only the electrical properties of the paint over time is observed, the paint is neither interactive nor adaptive but appears autonomous; indeed the properties change as a result of internal nondeterminism. But if that LoA is augmented by the structure data monitored by the paint, over time, then SmartPaint becomes an agent, because the data provide input to which the paint adapts its state. Finally, if that LoA is augmented further to include a model by which the paint works, changes in its electrical properties are revealed as being determined directly by input data and so SmartPaint no longer forms an agent.

#### 11.2.6.6 Organisations

A different kind of example of AA is provided by a company or management organisation. At an appropriate LoA, it interacts with its employees, constituent substructures and other organisations; it is able to make internally-determined changes of state; and it is able to adapt its strategies for decision making and hence for acting.

## 11.3 Morality

We have seen that given the appropriate LoA, humans, webbots and organisations can all be properly treated as agents. Our next task is to determine whether, and in what way, they might be correctly considered moral agents as well.

### *11.3.1 Morality of Agents*

Suppose we are analysing the behaviour of a population of entities through a video camera of a security system that gives us complete access to all the observables available at $LoA_1$ (see above 2.5) plus all the observables related to the degrees of interactivity, autonomy and adaptability shown by the systems under scrutiny. At this new $LoA_2$, we observe that two of the entities, call them H and W, are able:

(i) to respond to environmental stimuli—e.g. the presence of a patient in a hospital bed—by updating their states (interactivity), e.g. by recording some chosen variables concerning the patient's health. This presupposes that H and W are informed about the environment through some data-entry devices, for example some perceptors;
(ii) to change their states according to their own transition rules and in a self-governed way, independently of environmental stimuli (autonomy), e.g. by taking flexible decisions based on past and new information, which modify the environment temperature; and
(iii) to change according to the environment the transition rules by which their states are changed (adaptability), e.g. by modifying past procedures to take into account successful and unsuccessful treatments of patients.

H and W certainly qualify as agents, since we have only "upgraded" $LoA_1$ to $LoA_2$. Are they also moral agents? The question invites the elaboration of a criterion of identification. Here is a very moderate option:

(O) An action is said to be morally qualifiable if and only if it can cause moral good or evil. An agent is said to be a moral agent if and only if it is capable of morally qualifiable action.

Note that (O) is neither consequentialist nor intentionalist in nature. We are neither affirming nor denying that the specific evaluation of the morality of the agent might depend on the specific outcome of the agent's actions or on the agent's original intentions or principles. We shall return to this point in the next section.

Let us return to the question: are H and W moral agents? Because of (O), we cannot yet provide a definite answer unless H and W become involved in some moral action. So suppose that H kills the patient and W cures her. Their actions are moral actions. They both acted interactively, responding to the new situation with which they were dealing, on the basis of the information at their disposal. They both acted autonomously: they could have taken different courses of actions, and in fact we may assume that they changed their behaviour several times in the course of the action, on the basis of new available information. They both acted adaptably: they were not simply following orders or predetermined instructions. On the contrary, they both had the possibility of changing the general heuristics that led them to take the decisions they took, and we may assume that they did take advantage of the available opportunities to improve their general behaviour. The answer seems rather straightforward: yes, they are both moral agents. There is only one problem: one is a human being, the other is an artificial agent. The $LoA_2$ adopted allows both cases, so can you tell the difference? If you cannot, you will agree that the class of moral agents must include AAs like webbots. If you disagree, it may be so for several reasons, but only five of them seem to have some strength. I shall discuss four of them in the next section and leave the fifth to the conclusion.

### *11.3.2 A-Responsible Morality*

One may try to withstand the conclusion reached in the previous section by arguing that something crucial is missing in $LoA_2$. $LoA_2$ cannot be adequate precisely because if it were, then artificial agents (AAs) would count as moral agents, and this is unacceptable for at least one of the following reasons:

- *the teleological objection*: an AA has no goals;
- *the intentional objection*: an AA has no intentional states;
- *the freedom objection*: an AA is not free; and
- *the responsibility objection*: an AA cannot be held responsible for its actions.

#### 11.3.2.1 The Teleological Objection

The teleological objection can be disposed of immediately. For in principle $LoA_2$ could readily be (and often is) upgraded to include goal-oriented behaviour (Russell and Norvig 2010). Since AAs can exhibit (and upgrade their) goal-directed behaviours, the teleological variables cannot be what makes a positive difference between a human and an artificial agent. We could have added a teleological condition and both H and W could have satisfied it, leaving us none the wiser concerning their identity. So why not add one anyway? It is better not to overload the interface because a non-teleological level of analysis helps to understand issues in "distributed morality", involving groups, organizations institutions and so forth, that would otherwise remain unintelligible. This will become clearer in the conclusion.

#### 11.3.2.2 The Intentional Objection

The intentional objection argues that it is not enough to have an artificial agent behave teleologically. To be a moral agent, the AA must relate itself to its actions in some more profound way, involving meaning, wishing or wanting to act in a certain way, and being epistemically aware of its behaviour. Yet this is not accounted for in $LoA_2$, hence the confusion.

Unfortunately, intentional states are a nice but unnecessary condition for the occurrence of moral agenthood. First, the objection presupposes the availability of some sort of privileged access (a God's eye perspective from without, or some sort of Cartesian internal intuition from within) to the agent's mental or intentional states that, although possible in theory, cannot be easily guaranteed in practice. This is precisely why a clear and explicit indication is vital of the LoA at which one is analysing the system from without. It guarantees that one's analysis is truly based only on what is specified to be observable, and not on some psychological speculation. This phenomenological approach is a strength, not a weakness. It implies that agents (including human agents) should be evaluated as moral if they do play the "moral game". Whether they mean to play it, or they know that they are playing it, is relevant only at a second stage, when what we want to know is whether they are *morally responsible* for their moral actions. Yet this is a different matter, and we shall deal with it at the end of this section. Here, it is to sufficient to recall that, for a consequentialist, for example, human beings would still be regarded as moral agents (sources of increased or diminished welfare), even if viewed at a LoA at which they are reduced to mere zombies without goals, feelings, intelligence, knowledge or intentions.

#### 11.3.2.3 The Freedom Objection

The same holds true for the freedom objection and in general for any other objection based on some special internal states, enjoyed only by human and

perhaps super-human beings. The AAs are already free in the sense of being non-deterministic systems. This much is uncontroversial, scientifically sound and can be guaranteed about human beings as well. It is also sufficient for our purposes and saves us from the horrible prospect of having to enter into the thorny debate about the reasonableness of determinism, an infamous LoA-free zone of endless dispute. All one needs to do is to realise that the agents in question satisfy the usual practical counterfactual: they could have acted differently had they chosen differently, and they could have chosen differently because they are interactive, informed, autonomous and adaptive.

Once an agent's actions are morally qualifiable, it is unclear what more is required of that agent to count as an agent playing the moral game, that is, to qualify as a moral agent, even if unintentionally and unwittingly. Unless, as we have seen, what one really means, by talking about goals, intentions, freedom, cognitive states and so forth, is that an AA cannot be held responsible for its actions.

Now, responsibility, as we shall see better in a moment, means here that the agent, her behaviour and actions, are assessable in principle as praiseworthy or blameworthy, and they are often so not just intrinsically, but for some pedagogical, educational, social or religious end. This is the next objection.

#### 11.3.2.4 The Responsibility Objection

The objection based on the "lack of responsibility" is the only one with real strength. It can be immediately conceded that it would be ridiculous to praise or blame an AA for its behaviour, or charge it with a moral accusation. You do not scold your iPhone apps, that is obvious. So this objection strikes a reasonable note; but what is its real point and how much can one really gain by levelling it? Let me first clear the ground from two possible misunderstandings.

First, we need to be careful about the terminology, and the linguistic frame in general, used by the objection. The whole conceptual vocabulary of "responsibility" and its cognate terms is completely soaked with anthropocentrism. This is quite natural and understandable, but the fact can provide at most a heuristic hint, certainly not an argument. The anthropocentrism is justified by the fact that the vocabulary is geared to psychological and educational needs, when not to religious purposes. We praise and blame in view of behavioural purposes and perhaps a better life and afterlife. Yet this says nothing about whether an agent is the source of morally charged action. Consider the opposite case. Since AAs lack a psychological component, we do not blame AAs, for example, but, given the appropriate circumstances, we can rightly consider them sources of evils, and legitimately re-engineer them to make sure they no longer cause evil. We are not punishing them, anymore than one punishes a river when building higher banks to avoid a flood. But the fact that we do not "re-engineer" people does not say anything about the possibility of people acting in the same way as AAs, and it would not mean that for people "re-engineering" could be a rather nasty way of being punished.

Second, we need to be careful about what the objection really means. There are two main senses in which AA can fail to qualify as responsible. In one sense, we say that, if the agent failed to interact properly with the environment, for example, because it actually lacked sufficient information or had no alternative option, we should not hold an agent morally responsible for an action it has committed because this would be *morally unfair*. This sense is irrelevant here. $LoA_2$ indicates that AA are sufficiently interactive, autonomous and adaptive fairly to qualify as moral agents. In the second sense, we say that, given a certain description of the agent, we should not hold that agent morally responsible for an action it has committed because this would be *conceptually improper*. This sense is more fundamental than the other: if it is conceptually improper to treat AAs as moral agents, the question whether it may be morally fair to do so does not even arise. It is this more fundamental sense that is relevant here. The objection argues that AAs fail to qualify as moral agents because they are not morally responsible for their actions, since holding them responsible would be conceptually improper (not morally unfair). In other words, $LoA_2$ provides necessary but insufficient conditions. The proper LoA requires another condition, namely responsibility. This fourth condition finally enables us to distinguish between moral agents, who are necessarily human or super-human, and AAs, which remain mere efficient causes.

The point raised by the objection is that agents are moral agents only if they are responsible in the sense of being prescriptively assessable in principle. An agent *a* is a moral agent only if *a* can in principle be put on trial. Now that this much has been clarified, the immediate impression is that the "lack of responsibility" objection is merely confusing the *identification* of *a* as a moral agent with the *evaluation* of *a* as a morally responsible agent. Surely, the counter-argument goes, there is a difference between, on the one hand, being able to say who or what is the moral source or cause of (and hence it is accountable for) the moral action in question, and, on the other hand, being able to evaluate, prescriptively, whether and how far the moral source so identified is also morally responsible for that action, and hence deserves to be praised or blamed, and in case rewarded or punished accordingly.

Well, that immediate impression is actually mistaken. There is no confusion. Equating identification and evaluation is a shortcut. The objection is saying that identity (as a moral agent) without responsibility (as a moral agent) is empty, so we may as well save ourselves the bother of all these distinctions and speak only of morally responsible agents and moral agents as synonymous. But here lies the real mistake. We now see that the objection has finally shown its fundamental presupposition: that we should reduce all prescriptive discourse to responsibility analysis. Yet this is an unacceptable assumption, a juridical fallacy. There is plenty of room for prescriptive discourse that is independent of responsibility-assignment and hence requires a clear identification of moral agents. Good parents, for example, commonly engage in moral-evaluation practices when interacting with their children, even at an age when the latter are not yet responsible agents, and this is not only perfectly acceptable but something to be expected. This means that they identify them as moral sources of moral action, although, as moral agents, they are not yet subject to the process of moral evaluation.

If one considers children an exception, insofar as they are potentially responsible moral agents, another example, involving animals, may help. There is nothing wrong with identifying a dog as the source of a morally good action, hence as an agent playing a crucial role in a moral situation, and therefore as a moral agent. Search-and-rescue dogs are trained to track missing people. They often help save lives, for which they receive much praise and rewards from both their owners and the people they have located, yet this is not the relevant point. Emotionally, people may be very grateful to the animals, but for the dogs it is a game and they cannot be considered morally responsible for their actions. At the same time, the dogs are involved in a moral game as main players and we rightly identify them as moral agents that may cause good or evil.

All this should ring a bell. Trying to equate identification and evaluation is really just another way of shifting the ethical analysis from considering *a* as the moral agent/source of a first-order moral action *b* to considering *a* as a possible moral patient of a second-order moral action *c*, which is the moral evaluation of *a* as being morally responsible for *b*. This is a typical Kantian move, but there is clearly more to moral evaluation than just responsibility, because *a* is capable of moral action even if *a* cannot be (or is not yet) a morally responsible agent. A third example may help to clarify further the distinction.

Suppose an adult, human agent tries his best to avoid a morally evil action. Suppose that, despite all his efforts, he actually ends up committing that evil action. We would not consider that agent morally responsible for the outcome of his well-meant efforts. After all, Oedipus did try not to kill his father and did not mean to marry his mother. The tension between the lack of responsibility for the evil caused and the still present accountability for it (Oedipus remains the only source of that evil) is the definition of the tragic. Oedipus is a moral agent without responsibility. He blinds himself as a symbolic gesture against the knowledge of his inescapable state.

### *11.3.3 Morality Threshold*

Motivated by the discussion above, morality of an agent at a given LoA can now be defined in terms of a threshold function. More general definitions are possible but the following covers most examples, including all those considered in the present chapter.

A threshold function at a LoA is a function which, given values for all the observables in the LoA, returns another value. An agent at that LoA is deemed to be morally good if, for some pre-agreed value (called the tolerance), it maintains a relationship between the observables so that the value of the threshold function at any time does not exceed the tolerance.

For LoAs at which AAs are considered, the types of all observables can be mathematically determined, at least in principle. In such cases, the threshold function is also given by a formula; but the tolerance, though again determined,

is identified by human agents exercising ethical judgements. In that sense, it resembles the entropy ordering introduced in Floridi and Sanders (2001). Indeed the threshold function is derived from the level functions used there in order to define entropy orderings.

For non-artificial agents, like humans, we do not know whether all relevant observables can be mathematically determined. The opposing view is represented by followers and critics of the Hobbesian approach. The former argue that for a realistic LoA it is just a matter of time, until science is able to model a human as an automaton, or state-transition system, with scientifically determined states and transition rules; the latter object that such a model is in principle impossible. The truth is probably that, when considering moral agents, thresholds are in general only partially quantifiable and usually determined by various forms of consensus. Let us now review the examples from Sect. 11.2.6 from the viewpoint of morality.

#### 11.3.3.1 Examples

The futuristic thermostat is morally charged since the LoA includes patients' well-being. It would be regarded as morally good if and only if its output maintains the actual patients' well-being within an agreed tolerance of their desired well-being. Thus, in this case a threshold function consists of the distance (in some finite-dimensional real space) between the actual patients' well-being and their desired well-being.

Since we value our email, a webbot is morally charged. In Floridi and Sanders (2001) its action was deemed to be morally bad (an example of artificial evil) if it incorrectly filters any messages: if either it filters messages it should let pass, or allows to pass messages it should filter. Here we could use the same criterion to deem the webbot agent itself to be morally bad. However, in view of the continual adaptability offered by the bot, a more realistic criterion for moral good would be that at most a certain fixed percentage of incoming email be incorrectly filtered. In that case, the threshold function could consist of the percentage of incorrectly filtered messages.

The strategy-learning system Menace simply learns to play noughts and crosses. With a little contrivance it could be morally charged as follows.

Suppose that something like Menace is used to provide the game play in some computer game whose interface belies the simplicity of the underlying strategy and which invites the human player to pit his or her wit against the automated opponent. The software behaves unethically if and only if it loses a game after a sufficient learning period; for such behaviour would enable the human opponent to win too easily and might result in market failure of the game. That situation may be formalised using thresholds by defining, for a system having initial state $M$, $T(M)$ to denote the number of games required after which the system never loses. Experience and necessity would lead us to set a bound, $T_0(M)$, on such performance: an ethical system would respect it whilst an unethical one would exceed it. Thus the function $T_0(M)$ constitutes a threshold function in this case.

Organisations are nowadays expected to behave ethically. In non-quantitative form, the values they must demonstrate include: equal opportunity, financial stability, good working and holiday conditions toward their employees; good service and value to their customers and shareholders; and honesty, integrity, reliability to other companies. This recent trend adds support to our proposal to treat organisations themselves as agents and thereby to require them to behave ethically, and provides an example of threshold which, at least currently, is not quantified.

## 11.4 Information Ethics

What does our view of moral agenthood contribute to the field of information ethics (IE)? IE seeks to answer questions like: "What behaviour is acceptable in the infosphere?" and "Who is to be held morally accountable when unacceptable behaviour occurs?". It is the infosphere's novelty that makes those questions, so well understood in standard ethics, of greatly innovative interest; and it is its growing ubiquity that makes them so pressing.

The first question requires, in particular, an answer to "What in the infosphere has moral worth?". I have addressed the latter in Floridi (2003) and shall not return to the topic here. The second question invites us to consider the consequences of the answer provided in this chapter: any agent that causes good or evil is morally accountable for it.

Recall that moral accountability is a necessary but insufficient condition for moral responsibility. An agent is morally accountable for $x$ if the agent is the source of $x$ and $x$ is morally qualifiable (see definition O in Sect. 11.2.1). To be also morally responsible for $x$, the agent needs to show the right intentional states (recall the case of Oedipus). Turning to our question, the traditional view is that only software engineers—human programmers—can be held morally accountable, possibly because only humans can be held to exercise free will. Of course, this view is often perfectly appropriate. A more radical and extensive view is supported by the range of difficulties which in practice confronts the traditional view: software is largely constructed by teams; management decisions may be at least as important as programming decisions; requirements and specification documents play a large part in the resulting code; although the accuracy of code is dependent on those responsible for testing it, much software relies on "off the shelf" components whose provenance and validity may be uncertain; moreover, working software is the result of maintenance over its lifetime and so not just of its originators; finally, artificial agents are becoming increasingly autonomous. Many of these points are nicely made in Epstein (1997) and more recently in Wallach and Allen (2010). Such complications may lead to an organisation (perhaps itself an agent) being held accountable. Consider that automated tools are regularly employed in the development of much software; that the efficacy of software may depend on extra-functional features like interface, protocols and even data traffic; that software programs running on a system can interact in unforeseeable ways; that software may now be downloaded at the click of an icon in such a way that the user has no access to the code and its

| 1 | **General moral imperatives** |
|---|---|
| 1.1 | Contribute to society and human well-being |
| 1.2 | Avoid harm to others |
| 1.3 | Be honest and trustworthy |
| 1.4 | Be fair and take action not to discriminate |
| 1.5 | Honor property rights including copyrights and patents |
| 1.6 | Give proper credit for intellectual property |
| 1.7 | Respect the privacy of others |
| 1.8 | Honor confidentiality |
| 2 | **More specific professional responsibilities** |
| 2.1 | Strive to achieve the highest quality, effectiveness and dignity in both the process and products of professional work |
| 2.2 | Acquire and maintain professional competence |
| 2.3 | Know and respect existing laws pertaining to professional work |
| 2.4 | Accept and provide appropriate professional review |
| 2.5 | Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks |
| 2.6 | Honor contracts, agreements and assigned responsibilities |
| 2.7 | Improve public understanding of computing and its consequences |
| 2.8 | Access computing and communication resources only when authorised to do so |

**Fig. 11.3** The principles guiding ethical behaviour in the ACM code of ethics

provenance with the resulting execution of anonymous software; that software may be probabilistic (Motwani and Raghavan 1995); adaptive (Alpaydin 2010); or may be itself the result of a program (in the simplest case a compiler, but also genetic code, Mitchell 1998). All these matters pose insurmountable difficulties for the traditional, and now rather outdated view that one or more human individuals can always be found accountable for certain kinds of software and even hardware. Fortunately, the view of this chapter offers a solution—artificial agents are morally accountable as sources of good and evil—at the "cost" of expanding the definition of morally-charged agent.

### *11.4.1 Codes of Ethics*

Human morally-charged software engineers are bound by codes of ethics and undergo censorship for ethical and of course legal violations. Does the approach defended in this chapter make sense when the procedure it recommends is applied to morally accountable, AAs? Before considering the question ill-conceived, consider that the Federation Internationale des Echecs (FIDE) rates all chess players according to the same Elo System, regardless of their human or artificial nature. Should we be able to do something similar?

The ACM Code of Ethics and Professional Conduct, adopted by ACM Council on the 16th of October 1992 (http://www.acm.org/about/code-of-ethics) contains 24 imperatives, 16 of which provide guidelines for ethical behaviour (eight general and eight more specific; see Fig. 11.3), with further 6 organisational leadership imperatives, and 2 (meta) points concerning compliance with the Code.

Of the first eight, all make sense for artificial agents. Indeed, they might be expected to form part of the specification of any morally-charged agent. Similarly for the second eight, with the exception of the penultimate point: "improve public understanding". It is less clear how that might reasonably be expected of an arbitrary AA, but then it is also not clear that it is reasonable to expect it of a human software engineer. Note that wizards and similar programs with anthropomorphic interfaces—currently so popular—appear to make public use easier; and such a requirement could be imposed on any AA; but that is scarcely the same as improving understanding.

The final two points concerning compliance with the code (4.1: agreement to uphold and promote the code; 4.2: agreement that violation of the code is inconsistent with membership) make sense, though promotion does not appear to have been considered for current AAs any more than has the improvement of public understanding. The latter point presupposes some list of member agents from which agents found to be unethical would be struck.[3] This brings us to the censuring of AAs.

### *11.4.2 Censorship*

Human moral agents who break accepted conventions are censured in various ways, which vary from (a) mild social censure with the aim of changing and monitoring behaviour; to (b) isolation, with similar aims; to (c) capital punishment. What would be the consequences of our approach for artificial moral agents?

By seeking to preserve consistency between human and artificial moral agents, one is led to contemplate the following analogous steps for the censure of immoral artificial agents: (a) monitoring and modification (i.e. "maintenance"); (b) removal to a disconnected component of the infosphere; (c) annihilation from the infosphere (deletion without backup). The suggestion to deal directly with an agent, rather than seeking its "creator" (a concept which I have claimed need be neither appropriate nor even well defined) has led to a nonstandard but perfectly workable conclusion. Indeed it turns out that such a categorisation is not very far from that used by the standard anti-virus software. Though not adaptable at the obvious LoA, such programs are almost agent-like. They run autonomously and when they detect an infected file they usually offer several levels of censure, such as notification, repair, quarantine, deletion, with or without backup.

For humans, social organisations have had, over the centuries, to be formed for the enforcement of censorship (police, law courts, prisons, etc.). It may be that analogous organisations could sensibly be formed for AAs, and it is unfortunate that this might sound science fiction. Such social organisations became necessary with the increasing

[3] It is interesting to speculate on the mechanism by which that list is maintained. Perhaps by a human agent; perhaps by an AA composed of several people (a committee); or perhaps by a software agent.

level of complexity of human interactions and the growing lack of "immediacy". Perhaps that is the situation in which we are now beginning to find ourselves with the web; and perhaps it is time to consider agencies for the policing of AAs.

## 11.5 Conclusion

This chapter may be read as an investigation into the extent to which ethics is exclusively a human business. Somewhere between 16 and 21 years after birth, in most societies a human being is deemed to be an autonomous legal entity—an adult—responsible for his or her actions. Yet, an hour after birth, that is only a potentiality. Indeed, the law and society commonly treat children quite differently from adults on the grounds that not they but their guardians, typically parents, are *responsible* for their actions. Animal behaviour varies in exhibiting intelligence and social responsibility between the childlike and the adult, on the human scale, so that, on balance, animals are accorded at best the legal status of children and a somewhat diminished ethical status, in the case of guide dogs, dolphins, and other species. But there are exceptions. Some adults are deprived of (some of) their rights (criminals may not vote) on the grounds that they have demonstrated an inability to exercise responsible/ethical action. Some animals are held accountable for their actions and punished or killed if they err.

Into this context, we may consider other entities, including some kinds of organisations and artificial systems. I have offered some examples in the previous pages, with the goal of understanding better the conditions under which an agent may be held morally accountable.

A natural and immediate answer could have been: such accountability lies entirely in the human domain. Animals may sometimes appear to exhibit morally responsible behaviour, but lack the thing unique to humans which render humans (alone) morally responsible; end of story. Such an answer is worryingly dogmatic. Surely, more conceptual analysis is needed here: what has happened morally when a child is deemed to enter adulthood, or when an adult is deemed to have lost moral autonomy, or when an animal is deemed to hold it?

I have tried to convince the reader that we should add artificial agents (corporate or digital, for example) to the moral discourse. This has the advantage that all entities that populate the infosphere are analysed in non-anthropocentric terms; in other words, it has the advantage of offering a way to progress past the immediate and dogmatic answer mentioned above.

We have been able to make progress in the analysis of moral agenthood by using an important technique, the Method of Abstraction, designed to make rigorous the perspective from which the domain of discourse is approached. Since I have considered entities from the world around us, whose properties are vital to my analysis and conclusions, it is essential that we have been able to be precise about the LoA at which those entities have been considered. We have seen that changing the LoA may well change our observation of their behaviour and hence change the

conclusions we draw. Change the quality and quantity of information available on a particular system and you change the reasonable conclusions that should be drawn from its analysis.

In order to address all relevant entities, I have adopted a terminology that applies equally to all potential agents that populate our environments, from humans to robots, from animals to organisations, without prejudicing our conclusions. And in order to analyse their behaviour in a non-anthropocentric manner I have used the conceptual framework offered by state-transition systems. Thus the agents have been characterised abstractly, in terms of a state-transition system. I have concentrated largely on artificial agents and the extent to which ethics and accountability apply to them. Whether an entity forms an agent depends necessarily (though not sufficiently) on the LoA at which the entity is considered; there can be no absolute LoA-free form of identification. By abstracting that LoA, an entity may lose its agenthood by no longer satisfying the behaviour we associate with agents. However, for most entities there is no LoA at which they can be considered an agent. Of course. Otherwise one might be reduced to the absurdity of considering the moral accountability of the magnetic strip that holds a knife to the kitchen wall. Instead, for comparison, our techniques address the far more interesting question (Dennet 1997): "when HAL kills, who's to blame?". The analysis provided in the article enable us to conclude that HAL is accountable—though not responsible—if it meets the conditions defining agenthood.

The reader might recall that, in Sect. 11.3.1, I deferred the discussion of a final objection to our approach until the conclusion. The time has come to honour that promise.

Our opponent can still raise a final objection: suppose you are right, does this enlargement of the class of moral agents bring any real advantage? It should be clear why the answer is clearly affirmative. Morality is usually predicated upon responsibility. The use of LoA and thresholds enables one to distinguish between accountability and responsibility, and formalise both, thus further clarifying our ethical understanding. The better grasp of what it means for someone or something to be a moral agent brings with it a number of substantial advantages. We can avoid anthropocentric and anthropomorphic attitudes towards agenthood and rely on an ethical outlook not necessarily based on punishment and reward but on moral agenthood, accountability and censure. We are less likely to assign responsibility at any cost, forced by the necessity to identify a human moral agent. We can liberate technological development of AAs from being bound by the standard limiting view. We can stop the regress of looking for the *responsible* individual when something evil happens, since we are now ready to acknowledge that sometimes the moral source of evil or good can be different from an individual or group of humans. I have reminded the reader that this was a reasonable view in Greek philosophy. As a result, we should now be able to escape the dichotomy "responsibility + moral agency = prescriptive action" versus "no responsibility therefore no moral agency therefore no prescriptive action". Promoting normative action is perfectly reasonable even when there is no responsibility but only moral accountability and the capacity for moral action.

All this does not mean that the concept of "responsibility" is redundant. On the contrary, the previous analysis makes clear the need for a better grasp of the concept of responsibility itself, when the latter refers to the ontological commitments of creators of new AAs and environments. As I have argued elsewhere (Floridi and Sanders 2005; Floridi 2007), Information Ethics is an ethics addressed not just to "users" of the world but also to demiurges who are "divinely" responsible for its creation and well-being. It is an ethics of *creative stewardship*.

In the introduction, I warned the reader about the lack of balance between the two classes of agents and patients brought about by deep forms of environmental ethics that are not accompanied by an equally "deep" approach to agenthood. The position defended in this chapter supports a better equilibrium between the two classes *A* and *P*. It facilitates the discussion of the morality of agents not only in the infosphere but also in the biosphere—where animals can be considered moral agents without their having to display free will, emotions or mental states (see for example the debate between Rosenfeld 1995a; Dixon 1995; Rosenfeld 1995b)—and in what we have called contexts of "distributed morality", where social and legal agents can now qualify as moral agents. The great advantage is a better grasp of the moral discourse in non-human contexts. The only "cost" of a "mind-less morality" approach is the extension of the class of agents and moral agents to embrace AAs. It is a cost that is increasingly worth paying the more we move towards an advanced information society.

**Acknowledgement** This contribution is based on Floridi and Sanders (2004), Floridi (2008a, 2010a). I am grateful to Jeff Sanders for his permission to use our work.

## References

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*, 251–261.

Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed.). Cambridge, MA/London: MIT Press.

Arnold, A., & Plaice, J. (1994). *Finite transition systems: Semantics of communicating systems*. Paris/Hemel Hempstead: Masson/Prentice Hall.

Barandiaran, X. E., Paolo, E. D., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior – Animals, Animats, Software Agents, Robots, Adaptive Systems, 17*(5), 367–386.

Bedau, M. A. (1996). The nature of life. In M. A. Boden (Ed.), *The philosophy of life* (pp. 332–357). Oxford: Oxford University Press.

Cassirer, E. (1910). *Substanzbegriff Und Funktionsbegriff. Untersuchungen Über Die Grundfragen Der Erkenntniskritik.* Berlin: Bruno Cassirer. Trans. by Swabey, W. M., & Swabey, M. C. (1923). *Substance and function and Einstein's theory of relativity.* Chicago: Open Court.

Danielson, P. (1992). *Artificial morality: Virtuous robots for virtual games*. London/New York: Routledge.

Davidsson, P., & Johansson, S. J. (Eds.) (2005). Special issue on "on the metaphysics of agents". *ACM,* 1299–1300.

Dennet, D. (1997). When Hal kills, who's to blame? In D. Stork (Ed.), *Hal's legacy: 2001's computer as dream and reality* (pp. 351–365). Cambridge, MA: MIT Press.

Dixon, B. A. (1995). Response: Evil and the moral agency of animals. *Between the Species, 11*(1–2), 38–40.

Epstein, R. G. (1997). *The case of the killer robot: Stories about the professional, ethical, and societal dimensions of computing*. New York/Chichester: Wiley.
Floridi, L. (2003). On the intrinsic value of information objects and the infosphere. *Ethics and Information Technology, 4*(4), 287–304.
Floridi, L. (2006). Information technologies and the tragedy of the good will. *Ethics and Information Technology, 8*(4), 253–262.
Floridi, L. (2007). Global information ethics: The importance of being environmentally earnest. *International Journal of Technology and Human Interaction, 3*(3), 1–11.
Floridi, L. (2008a). Artificial intelligence's new frontier: Artificial companions and the fourth revolution. *Metaphilosophy, 39*(4/5), 651–655.
Floridi, L. (2008b). The method of levels of abstraction. *Minds and Machines, 18*(3), 303–329.
Floridi, L. (2010a). *Information – A very short introduction*. Oxford: Oxford University Press.
Floridi, L. (2010b). Levels of abstraction and the Turing test. *Kybernetes, 39*(3), 423–440.
Floridi, L. (2010c). Network ethics: Information and business ethics in a networked society. *Journal of Business Ethics, 90*(4), 649–659.
Floridi, L., & Sanders, J. W. (2001). Artificial evil and the foundation of computer ethics. *Ethics and Information Technology, 3*(1), 55–66.
Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.
Floridi, L., & Sanders, J. W. (2005). Internet ethics: The constructionist values of Homo Poieticus. In R. Cavalier (Ed.), *The impact of the internet on our moral lives*. New York: SUNY.
Franklin, S., & Graesser, A. (1997). Is it an agent, or just a program?: A taxonomy for autonomous agents. In *Proceedings of the workshop on intelligent agents III, agent theories, architectures, and languages* (pp. 21–35). Berlin: Springer.
Jamieson, D. (2008). *Ethics and the environment: An introduction*. Cambridge: Cambridge University Press.
Kerr, P. (1996). *The grid*. New York: Warner Books.
Michie, D. (1961). Trial and error. In A. Garratt (Ed.), *Penguin science surveys* (pp. 129–145). Harmondsworth: Penguin.
Mitchell, M. (1998). *An introduction to genetic algorithms*. Cambridge, MA/London: MIT.
Moor, J. H. (2001). The status and future of the Turing test. *Minds and Machines, 11*(1), 77–93.
Motwani, R., & Raghavan, P. (1995). *Randomized algorithms*. Cambridge: Cambridge University Press.
Moya, L. J., & Tolk, A. (Eds.). (2007). Special issue on towards a taxonomy of agents and multi-agent systems. *Society for Computer Simulation International,* 11–18.
Rosenfeld, R. (1995a). Can animals be evil?: Kekes' character-morality, the hard reaction to evil, and animals. *Between the Species, 11*(1–2), 33–38.
Rosenfeld, R. (1995b). Reply. *Between the Species, 11*(1–2), 40–41.
Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd, International). Boston/London: Pearson.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.
Wallach, W., & Allen, C. (2010). *Moral machines: Teaching robots right from wrong*. New York/Oxford: Oxford University Press.

# Chapter 12
# The Good, the Bad, the Ugly… and the Poor: Instrumental and Non-instrumental Value of Artefacts

**Maarten Franssen**

**Abstract** Technical artefacts are subject to normative judgements, in particular evaluative judgements, as a matter of course: we speak of good saws, poor drills, and so forth. These judgements concern the instrumental value of artefacts: a saw is good as a saw, a drill is poor as a drill. In this essay I investigate whether we can also attribute non-instrumental value to artefacts, where we would judge an artefact to be good or bad not in the sense of being an instrumentally good saw or poor drill but being a morally good saw or bad drill. Adopting an overall view of normativity that takes reasons for action or thought as the fundamental notion and that links the value of anything that has value to the existence of reasons to create or promote it in case of positive value or goodness and to the existence of reasons to eliminate or fight it in case of negative value or badness, I defend a view that artefacts can be evaluated as bad or good not on the basis of how they are used but on the basis of their design. Additionally I look into the question whether this analysis applies to an equally extent to judgements of artefacts as bad and judgements of artefacts as good and show some form of asymmetry between the two. Finally I extend the analysis beyond the class of technical artefacts to moral judgements of other artefacts, notably works of art.

## 12.1 Introduction: Agency, Normativity and Values

In this chapter, I will adopt a traditional approach to the connection between agency and artefacts: I will assume that agency emerges from intentional agents and is directed towards artefacts, as objects that intentional agents do something about or with. Although various animals, in particular birds and mammals, may show some

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

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_12,

form of intentionality, I will also adopt the traditional point of view that normal, adult people are currently the only beings known to us to which a description as intentional agents fully applies, meaning that people have beliefs and expectations about the world as it is, have desires and goals concerning the world as it could be, and act on the basis of reasons they perceive themselves to have, given their beliefs and goals. I will not give arguments here why one should adopt either of these two traditional positions and why I find none of the arguments around for ascribing a form of agency to artefacts convincing, since this essay is about other issues. I simply state this view as my point of departure.

A crucial aspect of agency is its openness to normative judgements. Our actions are open to reflection and to critique as being right or wrong, whether it be prior to acting or after the act, and whether it be within our own mind or as a form of public scrutiny. Still, the scope of such normative judgements seems wider than just our actions. In particular artefacts seem to participate in this practice of conferring normative judgements: we routinely speak of good and poor cars, good and poor drills, and so on. In line with the previous paragraph, however, I take it that all normative judgements are grounded in the character of human agency, and that the judgements of artefacts as good or poor must be understood as being grounded in this way. In previous work on the relation of artefacts and normativity I have analysed evaluative judgements about technical artefacts, stating the goodness or poorness, as expressing their instrumental value and have argued how the normative content of such judgements derives from the way they figure in human action.[1] In this essay I inquire whether a similar approach can ground the attribution of non-instrumental, moral value to artefacts.

In my analysis of normative statements involving artefacts I follow Dancy (2006) in characterizing normativity as referring to a relation that the facts of the world bear to our question as human beings what to do and what to believe, given that we are – or at least experience ourselves to be – freely acting beings continuously faced with a variety of options for thought and action. On this view, the key primitive term of the vocabulary of normativity is 'reason'. Certain facts about the world give us reasons for or against doing something (irrespective of whether we perceive these facts to be such reasons). On the one hand this grounds the deontic; an action is right if the balance of all reasons for or against it swings towards performing that action, or, more strongly, if there is a compelling or conclusive reason to perform it, and it is wrong if the balance of reasons favours not performing it or if there is a compelling reason not to perform that action. On the other hand, it grounds our attributions of value, of distributing the evaluative labels 'good' and 'bad'.

In the closest-knit view of the normative, states-of-affairs are the primary 'entities' that have value, i.e. are good or bad, since by our actions we change existing states-of-affairs and bring about new ones. A state-of-affairs, then, is good if we have reason to bring it about, or to maintain or sustain it if it obtains, and a state-of-affairs is bad if we have reason to bring it to an end, or to change it or oppose it if it obtains. In this way the rightness of an action is directly connected to

[1] Franssen (2006, 2009).

the value of the resulting state-of-affairs: an action is right because one has a conclusive to bring about the state-of-affairs that will result from the action, or the balance of reasons swings toward bringing about that state-of-affairs. And so, *mutatis mutandis*, is the wrongness of an action.[2]

This may, however, be too restricted an understanding of the way the deontic and evaluative aspects of normativity hang together. It may be that we can have reasons for actions that are not exhausted by the reasons we have for bringing about the resulting state-of-affairs. And it seems we also value more than just states-of-affairs. For once, we value abstract 'ideals', which are typically referred to themselves as *values*, such as friendship, justice, privacy, and so forth. To value these is not to have reasons to bring it about that they exist, although, arguably, it may imply reasons with respect to the bringing about of manifestations of them. To account for such 'valuings', Scanlon (1998) and Dancy (2000) have argued for a more general account, which holds that something being good means that one has reasons for taking a broadly positive action or adopting a broadly positive attitude with respect to it – such as admiring it, protecting it, promoting it, cherishing it, and so forth – and something being bad means that one has reasons for taking a broadly negative action or adopting a broadly negative attitude with respect to it – such as holding it in contempt, rejecting it, avoiding it, and the like. There is a threat that this trivializes the notion of value, because what it is exactly that we have reasons for or against becomes vague and arbitrary, and also a threat of circularity or regress, because good and bad are traded in for roughly positive and negative attitudes, and what then grounds this distinction? In judging the seriousness of these difficulties, however, it should be emphasized that the account aims to clarify what sort of judgements normative judgements are, what the relations between the kinds of normative judgements are and the key concepts occurring in them, and not why we hold some of them true and others false, or accept some of them and reject others. I will not further discuss such difficulties in this paper, but adopt the perspective sketched to discuss the different forms of value attributed to artefacts both from a broad view of goodness (badness) – where goodness (badness) is having properties that give reasons for a positive (negative) attitude – and from a narrow view – where goodness (badness) is having properties that give reasons for (against) bringing about or bringing into existence.

On the view of normativity adopted from Dancy, normative aspects of states-of-affairs and things can be seen as second-order aspects, because they explicate how their first-order aspects, their direct, descriptive attributes, give us reasons for certain actions. The value of something, its goodness or badness, is not a property of it on a par with its first-order properties such as its physical or causal properties. States-of-affairs and things have their first-order aspects irrespective of the existence of intentional beings but their second-order aspects only in relation to these intentional beings. Again, it may remain a point of discussion what exactly is to be

[2] This ignores the aspect of uncertainty, that is, that we typically cannot be sure which state of affairs will result from an action. This aspect is central to formal theories of action, such as the theory of rational choice. I will, however, ignore this aspect in this paper.

included in the totality of first-order properties that give us reasons, and accordingly whether the distinction between first-order properties and second-order properties is clear-cut in the first place. However, this issue also lies beyond this paper. The central issue here is to see how this account helps us in making sense of the sort of things people say concerning the values of objects.

As stated, it is perfectly accepted to speak of a good instrument, say, a good drill, and this indicates that, next to states-of-affairs or 'ideals', we also attribute what is *prima facie* a form of value to objects. At least for those objects that are artefacts, this attribution of value can be understood in accordance with the approach to normativity adopted here as indicating the presence of properties that are reason-giving, and elsewhere I have elaborated in detail how this can be done (Franssen 2006, 2009). To call a particular technical artefact a good specimen of its functional kind is to express the fact that it has properties that give someone who has a reasonable need for achieving the purpose for which this kind of artefact was designed a reason to use it for this purpose. For example, a good knife is a knife that has properties that give someone who has a reasonable need to cut something a reason to use this particular knife for cutting. The goodness of the knife is a second-order attribute with respect to, say, its sharpness, to settle on just one relevant property: it is the first-order aspect of its sharpness that gives us a reason to use the knife for cutting with, and this being so is a second-order aspect that is expressed by calling the knife a good knife. In this way attributions of *instrumental* value can be shown to be a particular sort of normative judgements.

However, on the reason-giving account of normativity and the linkage of positive value to reasons for bringing about or the adoption of a broadly positive attitude and of negative value to reasons for ending or the adoption of a broadly negative attitude, this analysis of instrumental value seems to make it a special sort of value, almost a deficient sort of value. The reasons that the properties of instrumentally good artefacts give us are reasons to *use* these artefacts, but using is not in any way either a positive or a negative sort of action. And certainly using an artefact is not a form of bringing it into existence: using an artefact presupposes its existence. Neither can using an artefact be connected to an *attitude* one adopts with respect to it. Properties of a knife like its sharpness are relevant for the specific question what to do *with* it, but do not bear direct relevance to the more general question what to do *concerning* it. The account of normative judgements adopted here suggests the latter as the decisive question with respect to it. Ross and Dancy seem to have had this aspect of instrumental value in mind in their statements that instrumental value is not a form of value at all.[3]

This analysis thus brings to light a clear difference between the character of value attributions to states-of-affairs and of attributions of instrumental value to instruments and other sorts of technical artefacts. Instead of 'This is a good state-of-affairs' we can also say 'This is a state-of-affairs and it is good', whereas instead of 'This is a good instrument' we cannot say 'This is an instrument and it

[3] Dancy (2000, p. 159).

is good'. In the latter case we cannot detach the instrument's goodness from its being an instrument: the artefact is not just good, it is good as an instrument, of a particular sort. A good saw is not a good knife. There is a difference, then, between instrumental value and non-instrumental value, or value 'as such', or moral value.[4]

This result in no way closes the door on the possibility of non-instrumental value judgements concerning technical artefacts, or even concerning a wider class of objects, or any object. Indeed we are inclined to make such judgements, in particular judgements of negative value, that is, judgements of instruments as bad in a moral sense. Typically, they are weapons: atomic bombs, particularly nasty land mines, poison gas, instruments of torture. My personal favourite as the quintessential bad artefact is the machine that is almost the main character in Kafka's story 'In the penal colony', which grafts an account of the (mis)deeds into the skin of the convicted. The badness of such technical artefacts is definitely not instrumental poorness. Kafka's torturing instrument is as bad – one would even say evil – as it is exactly because as an instrument of torture, it is quite good, that is, performs well. We acknowledge that the thing is an (instrumentally) good instrument of torture. Precisely because of that, what we want to say is that the thing is an instrument of torture and is bad or evil.

In this essay, then, I investigate whether the general account of normativity that puts reasons central can account for the attribution of non-instrumental value to artefacts, and I show how it can. A consequence of the analysis is that such attributions are in order only for human-made objects, not for objects in general, but this includes other objects apart from technical artefacts, e.g. artworks. By arriving in this way at a unified account of value – instrumental and non-instrumental, foremost but perhaps not exclusively moral, value – attributed to artefacts, we also have an explanation of why the words 'good' and 'bad' are used both for the instrumental value of objects and the moral value of states-of-affairs. (In English, for the opposite of 'good' in the instrumental sense, 'poor' is preferred, but other languages do not make this distinction, and even in English, we would say that a particular stone is a bad one to use as a hammer next to saying that it makes a poor hammer.) Additionally, the account can be tested for the extent to which it supports an apparent asymmetry between judgements of artefacts as bad and judgement of them as good. Intuitive examples of quintessentially bad artefacts come to mind easily, but equally intuitive examples of quintessentially good artefacts do not. I argue, however, that the account allows for only a limited asymmetry between good and bad. The account can also be tested by the extent to which it allows for the attribution of non-instrumental value to non-technical artefacts, in particular artworks; this is taken up in the final section.

[4] One may judge that it requires further argumentation that 'value as such', the value that a state-of-affairs has such that one has reasons to bring it about, is moral value. I will not do so here, however, and, for reasons of simplicity, equate the notions of non-instrumental value and moral value.

## 12.2 A Preliminary: Individual Objects vs. Artefact Kinds

In investigating the various sorts of values artefacts can have, the focus is on their being representatives of *artefacts kinds* – i.e., knife, gun, bomb, torturing device –, not mere individual objects.[5] We may have 'personal' reasons for specific pro-attitudes with respect to particular artefactual objects, that is, take ourselves to have a reason for such pro-attitudes. I may take it that I have such reasons, say, with respect to the medal that saved my great-grandfather's life in the trenches by deflecting a bullet that would otherwise have killed him. But what reason exactly, and for what pro-attitude? I do not have reason to wear it as a medal; that would even be cheating, since as a medal it was not awarded to me. As another example, take the parachute that saved my life once. I do not now have more reason to use this parachute rather than another parachute when I make my next jump. Presumably I have less reason, because it is by now an old parachute and there will be risks involved in using it that are lacking when I use a 'fresh' parachute. If we take such artefact tokens to have a special value for us, it is not an increased instrumental value, which give additional reasons for using them.

There may be examples that point in the opposite direction. I may take myself to have a reason to use the fountain pen that belonged to my grandfather for writing, rather than another pen, or to use, say, the fountain pen, bought at an auction, that once belonged to Thomas Hardy, whose novels I much admire. But such considerations seem still not to be related to its instrumental value as a pen. It may even be more difficult for me to write with such a pen than with a modern one. So the reasons at work here are different ones from instrumental reasons: I cherish the remembrance of my grandfather, or express my admiration for Hardy, through using the pen, and cherishing or expressing admiration are the relevant actions, not writing. Only if I hope to write better fiction by using Hardy's pen would my reason be an instrumental sort of reason, but this is hardly a convincing reason.

Such reasons with respect to individual objects therefore do not ground their instrumental value, since it is not generally the case that someone with the corresponding goal has a reason to use them. Nor do they ground moral value in the sense that we are after here. The reasons I may take myself to have for cherishing the medal or the parachute, caring for it, or whatever pro-attitude seems in order in a particular case, are not generalizable to other people. No-one else would see a reason to adopt a similar pro-attitude toward my grandfather's medal, nor would I claim that they should adopt a similar pro-attitude. Of course I will claim that others have reasons for particular attitudes towards it, but that is because they should take my valuations into account in handling certain objects. If someone damages the medal or is careless about it I will blame that person, but not because that person

[5] I have elsewhere emphasized a distinction between artefact kinds as primarily functional kinds, defined by a particular form of use, like 'knife', and artefact kinds as primarily structural kinds, defined by minimally an operational principle and possibly further details about it make-up. This distinction will matter only in Sect. 12.4 and I will not elaborate it here; see note 7 for more details.

fails to copy my pro-attitudes towards the medal – that of cherishing it – but because that person hinders my cherishing it, without there being a good reason for doing so.

What is more, such personal valuations are often independent of the valued objects being artefacts; they could be natural objects just as well. The bullet that failed to kill my great-grandfather could have been deflected by a pebble rather than a medal, and I could cherish Hardy's shell used as a paperweight rather than his pen. So we can learn little about either the instrumental or non-instrumental value of artefacts from such cases.

To be sure, there are cases where the value of token artefacts is not limited to individual people, because they have value for groups of people, as symbols. When Nazi Germany had defeated France in June 1940, Hitler insisted that the capitulation of the French army be signed in the same rail carriage in which the German army representatives had signed the armistice in November 1918 – the French had obligingly kept it a museum – and at the very same spot, the wood of Compiègne. We understand why Hitler arranged things this way, and so did most Germans, meaning that they would have said that anyone in Hitler's position would have had a reason to do the same thing. Virtually anything could be valuable in this way, depending on the precise history that clings to it. This is still not what we are after when discussing whether artefacts have non-instrumental values next to their instrumental value. In the Compiègne case, it is more in order to say that this particular railway carriage was valuable, both to the French and to the Germans, but typically with opposing values for different groups. It does not imply anything about the instrumental value of railway carriages beyond what its use in 1918 had already taught us – that they can be used to hold meetings in – nor does it imply anything about the positive or negative moral value of railways carriages as such, in the sense in which saying that Kafka's torturing machine is bad extends to any specimen of it that would exist anywhere.

The issue, then, is whether artefacts can have non-instrumental value next to their instrumental value as representatives of artefact *kinds*, like 'railway carriage'. This issue concerns artefact kinds instead of individual artefactual objects in a stricter way than the question of the instrumental value of artefacts does. A broken drill or a worn and blunt knife are poor instruments as individual objects, even if they once started out as instances of a good kind of drill or knife. In posing the question whether a drill or a knife could be termed a good or bad artefact, the fact that drills and knives come in fresh and worn specimens, instrumentally good and poor copies, is not directly relevant.

## 12.3 The Badness of Artefacts Grounded in Their Use?

So then what does make artefacts, as representatives of artefact kinds, good or bad in a non-instrumental, moral sense? Because I will argue that there is to some extent an asymmetry between good and bad with respect to artefacts, let us first concentrate on an analysis of what can make artefacts bad. Von Wright, in his important but

somewhat neglected work on value (1963), already pointed out the distinction between instrumental value and moral value, reflected in English by the distinction between a poor instrument and a bad instrument. However, he only cursorily addressed the issue of moral value of artefacts and things in general, dealing with it in one sentence: "An artefact is bad if it has detrimental side-effects." This grounds the instrumental value of an artefact in its intended consequences and the moral value of an artefact in its non-intended consequences. The suggestion raises a great many questions, however, in particular with respect to how the notion of 'side-effects' is to be understood. Is an artefact bad if it has detrimental side-effects when used only once, or when used regularly? Do these side-effects have to result right away or in the long run? Has the side-effect to be necessary or is contingent sufficient? Must the side-effect result from using the artefact according to its proper function or is it allowed that it was used according to an accidental function? Should the side-effect have been foreseeable or is that unimportant?

If we side-step all these questions and assume that the detrimental side-effects must occur with any single use, not many artefacts inevitably have such side-effects. Perhaps only nuclear weapons qualify, although it is not obvious what should be termed their detrimental effects and their detrimental side-effects. Perhaps the radioactive fallout after the explosion must be considered a side-effect. But even in the case of nuclear weapons, these side-effects are detrimental only when used as weapons in inhabited areas. When used for massive-scale construction works in deserted areas, or to destroy a comet heading for the earth, the side-effects are no longer obviously detrimental. The resulting radioactive fallout is harmful to some extent, but 'detrimental', I would say, refers to something more serious. The contribution to radiation levels on a global scale is easily comparable to forms of pollution resulting from the massive use of any form of technology. In this respect, the accumulated side-effects of the joint use of technical artefacts has not turned out detrimental yet, by the timely taking of measures, but they could in the future, for example in the case of global warming. However, we cannot lead back these conditions to the effects of the use of one particular artefact or artefact kind, for example cars with an internal-combustion engine. One can only evaluate the totality of technology as bad on the basis of detrimental side-effects (assuming they will one day prove to be truly and uncontroversially detrimental). But this will be evaluating a form of life as bad, not a particular artefact or artefact kind.

Langdon Winner has argued in a more subtle way in his (1986) that the technology of nuclear energy is bad, and thereby presumably also its central artefacts, like nuclear power plants, because of its side-effects. According to Winner, the adoption of nuclear energy necessitates a regime of tight control, which is likely to foster a development of society away from democracy. This, however, is too speculative to be convincing. Why should strict safety and security regimes be incompatible with democracy, or be conducive to developments away from democracy? If such things occur, it will hardly be the implementation of the safety regime in isolation that caused such a development. Again, the criticism concerns a form of life, not a particular kind of artefact; the argument that the artefact inevitably leads to the form of life and can therefore be equated with it as far as moral judgement is concerned is not convincing.

The basic objection against von Wright's proposal is that it misses the point. If the goodness or badness of technical artefacts lies in their effects, then the quintessential candidates for bad artefacts – pernicious weapons, torturing instruments, Kafka's machine in the penal colony – are not bad because they have detrimental side-effects but because they have detrimental effects. Instruments that have detrimental side-effects could be considered poor instruments, open for redesign until the detrimental side-effects no longer occur or are no longer detrimental.[6] The quintessential examples, however, combine badness and instrumental goodness, and are the more bad for being instrumentally good.

This fits into the reason-giving account of the normative adopted here. The fact that using an instrument in a particular way creates detrimental side-effects is overwhelmingly reason-giving by presenting reasons *not to use* the instrument *in that way*. This makes the aspect of detrimental side-effects part of the instrumental value of an artefact. They add to reasons for and against using it. The mere existence of detrimental side-effects cannot count as a sufficient reason against creating the artefact, for all sorts of reasons: there may be other forms of use that avoid the detrimental side-effects, or these side-effects can be contained, or the beneficial results of their use are considered, at least occasionally, to outweigh the side-effects, or the artefact can be developed further so as to make the side-effects less detrimental, and so on.

This suggests a modification of von Wright's proposal, which would deliver a more satisfactory characterization: an artefact is bad not if it has detrimental side-effects but if it has detrimental effects, or in other words, if it is used overwhelmingly in order to realize detrimental outcomes. Kafka's quintessentially bad device certainly would qualify as bad on this account. The difficulty, however, is that this criterion qualifies far too many artefacts as bad. Suppose, for example, that all criminal violence were done with guns, and that guns were also the one weapon that is available to the police to battle crime. Assume that the police would prefer not to shoot their guns, and if they have to, to use them to incapacitate, not to kill, in contrast to the way criminals use guns. As a result, the overwhelming use of guns would be their criminal use. But we would not want to say that this makes guns bad, whether in the hands of the police or of criminals. There would be something perverse in claiming that the moral status of the guns that are the police's only recourse against crime is determined by the use criminals make of it.

In general, for most artefacts, even if they not only can be used to realize bad outcomes but in fact are typically or overwhelmingly used to achieve bad outcomes, there are still ways of using them in order to realize good outcomes. Therefore, it is not generally true that they should not be used, i.e. that there are compelling reasons not to use them, or the balance of reasons favours not using them. Nor is there a compelling reason not to produce them. Although it were perhaps preferable if no guns or weapons existed at all, given that they do it is

[6] Indeed Hansson (2006), following Godlovich, includes the absence of undesired (side-)effects in the functional characteristics of technical artefacts.

reasonable to produce enough of them to equip the police with them in order to enable the police to fight and contain crime.

Such considerations, however, seem not to apply to Kafka's torturing machine. Perhaps, then, the class of bad artefacts could be said to consist of those artefacts for which exclusively a form of use resulting in bad outcomes seems possible. There could be weapons, for example, that should not be used to fight criminals or fight a war for just causes, even if the criminals or the enemy are using them. This consideration points in the right direction, I think, but still the criterion should not be articulated in terms of the forms of use available for such artefacts. It would make the class of bad artefacts depend too much on our (lack of) imagination. Perhaps some day a more innocent use is discovered of landmines that, when triggered, jump up about one meter into the air before exploding, so as to create wounds that are messy and difficult to treat and in that way not only incapacitate soldiers or fighters but also clog the adversary's medical facilities and destroy its morale. Perhaps some day even a more innocent use for Kafka's machine is discovered. To concentrate on the possibilities for use of quintessentially bad artefacts is somehow to miss what makes them bad. In the next section I argue for an alternative view, which holds that a bad artefact is bad by design.

## 12.4 The Badness of Artefacts Is Badness by Design

My central claim in this paper is that the relevant action that is the focus of reasons in the case of bad artefacts, or good artefacts for that matter, is the action of *designing* them, not using them. Bringing them into existence is precisely what we do with artefacts: their existence is our responsibility. An artefact is bad if designing it is wrong, meaning that it should not be brought into existence, or that there is a compelling reason against bringing it into existence. The considerations in the previous section make clear that this compelling reason is not the fact that its existence merely allows its being used in a bad way or for a bad purpose. One can use almost any artefact for a bad purpose; as a consequence almost any artefact would be bad. The reason that one should not design such an artefact is that it is *explicitly meant* to be used for bad purposes. In my view, this excludes most ordinary weapons from being bad, since although they may often be used for bad purposes, it cannot be defended that they are meant to be used for bad purposes primarily. One can imagine designing a gun with the sole intention of making it available for the protection of its user from attacks by criminals or wild animals, that is, as a weapon of self-defence.

This cannot apply to instruments of torture. An instrument of torture is designed to inflict great pain to human persons. But what makes it so? It cannot be read off of the physical device itself. What an artefact's function could possibly be is at most constrained by its physical make-up. I take an artefact to be an object designed and made for a specific purpose. The *kind* of artefact a particular device belongs to is jointly determined by the functional requirements and the design specifications. The

functional requirements contain the details on what the artefact must be able to achieve, or allow us to achieve by using or implementing it, whereas the design specifications fix the physical route to the realization of this achievement. Artefacts can therefore be classified into kinds in two ways: into functional kinds, by the mere specification of (some of) the functional requirements, or into more narrowly defined artefact kinds, which add details about the operational principle and physical mechanism realizing the artefact's function(s).[7] It is the definition of the design task that stood at the beginning of the design and manufacture process resulting in copies of the artefact that determines an artefact's badness. It presents any designer with a compelling reason not to bring such an artefact into existence, and bringing it into existence would be wrong.

This characterization can deal with some of the excuses that people who are involved in the designing of bad artefacts may come up with – the designer's 'Ich habe es nicht gewußt'. One cannot be excused from designing a bad artefact by 'identifying' the artefact through its design specifications or its blueprint and claiming that one did not know its intended users and therefore could not know that they meant bad with it. It is not so much the artefact itself as the artefact's functional requirements that proclaim its badness, by specifying, for instance, that the design task is aimed at the creation of an instrument of torture. Functional requirements, however, can be phrased in various ways, and this opens up an escape route for designers who seek an excuse from having been or being involved in the design of a bad artefact. It seems feasible to make any design task innocuous by rephrasing the requirements such that all references to the damage they are intended to produce in use is couched in neutral language. This may not be an easy task for torturing instruments, and not a task I am eager to undertake, but one can imagine how it could be done for the example given in the previous section of landmines that jump up about one meter into the air before exploding, or for landmines producing shrapnel of plastic that is difficult to detect with x-rays. The functional requirements for such landmines would merely require them to be propelled to a certain elevation before exploding, or to be made of a material that is not metallic but still tears with sharp edges. These requirements do not broadcast the resulting artefact's badness. To deal with that, it is necessary to adopt the proper perspective at what defines a design task.

Most analyses of the design process carve it up into several consecutive phases. At the start are what are usually called customer needs; they represent what the ultimate user intends to achieve with the aid of the artefact. These will then be translated into the functional requirements: the totality of unambiguously defined behaviour or characteristics the artefact will have to be able to show. These again will have to be translated into the design specifications: the precise physical properties – material,

[7] See (Franssen 2009) and (Franssen 2013) for more details about the distinction between functional kinds and artefact kinds. The term 'artefact kind' for the more narrowly defined kind is my own; the term 'functional kind' is of course widely used, but seldom with the understanding that it is just one way among several of classifying artefacts (or other items to which functions are attributed, such as biological items) into kinds.

geometry – of a device such that if it has these properties, it will show the characteristics or be capable of the behaviour specified in the functional requirements. Part of the design task is to see to it that the customer's needs, typically articulated vaguely and fragmentarily in non-technical language, have been translated correctly, i.e., in line with the customer's intentions, into functional requirements that are delineated sharply enough to allow the setting of the design specifications. If necessary, it belongs to the responsibility of the designer to check that the 'needs' of the customer have been understood correctly. If the full extent of a design process is taken into account, it is clear that a designer cannot hide behind an projected artefact's functional requirements. It belongs to the responsibility of a designer to understand why it is expected of a landmine under design that it is propelled one meter into the air before exploding, or that its non-metal casing must fracture into bits with sharp edges. A designer must know, for example, whether an elevation of just half a meter or up to two meters would still do.[8]

In reverse, knowledge of the motivation of a particular artefact's functional requirements is what exculpates the design of good artefacts whose functional requirements may seem suspect. Take as one of the requirements for a particular knife that it should cut easily through human skin without too much pressure being required. If presented with this design task, it is important to know that what has to be designed is a medical scalpel.

In bringing in the 'needs' of the customer for judging the rightness or wrongness of creating a particular artefact, care must be taken not to extend this too far. Intentions are relevant only to the extent that they are reflected in the functional requirements 'defining' the artefact. Take, for example, the decision by the United Kingdom and the United States to develop the atomic bomb during the second world war: it can be judged, for all the dreadfulness of the bomb as a weapon, as morally justified by the circumstances, as the least of two evils, and forced upon these countries by conditions having the general aspect of a prisoner's dilemma. The decision by Nazi-Germany to develop an atomic bomb, if it had actually been taken, would have been wrong, for the same reasons that many decisions taken by Nazi-Germany were morally wrong, because they established and consolidated one of the most murderous regimes the world has ever seen. The decision of the UK and the USA to develop it given their justified opinion that in Nazi-German there were plenty of people who were aware of the possibility of such a weapon and a sufficient number of people who had the knowledge and resources necessary for actually developing one, was, arguably, right because it was directed at putting an end to this murderous regime.

[8] My analysis supposes that the instrument emerging from the design actually is capable of performing as intended to a sufficient extent. An instrument of torture that was designed so grossly inadequately that one could not possibly use it to inflict pain on anyone and would even find positive applications once its failure as an instrument of torture had become obvious could hardly be termed a bad artefact and would therefore pose a counter-example to the present analysis. However, most accounts of technical artefacts contain some form of success condition for the realization of the designer's intentions for an item to be a member of a particular artefact kind; e.g. Thomasson (2003).

This cannot be interpreted as implying that the allied atomic bomb was a morally good artefact whereas an atomic bomb developed by Nazi-Germany would have been a morally bad artefact. To be sure, artefact kinds are identified by their intentional history: artefacts are things designed for a particular purpose. As already stated, from the mere physical make-up of a device one cannot derive its proper function, and therefore not what kind of artefact it is, or even whether it is an artefact in the first place. By bringing in additional elements of an artefact's history we could introduce finer types. Obviously there is a limit to this: not all details from an artefact's history can be accepted as relevant to the definition of *artefact types*. However, the decisions that an artefact's designer must make on how to interpret the functional requirements and there relative importance for the sake of making trade-offs can be considered relevant in this respect, such that different decisions lead to different types. Still, the design tasks for the allied and Nazi atomic bombs, had the Nazi effort actually developed as far as the articulation of a design task, can be assumed to be roughly identical: to cause an explosive chain reaction of splitting atoms in a piece of material. What either party aimed to achieve by realizing such explosions was not reflected in this task in the allied case or would not have been in the Nazi case.[9]

Intentions, then, to the extent that they inform functional requirements, enter into a judgement concerning the goodness or badness, but these intentions should be distinguished both from the intentions guiding the decision to design in the first place and any decision to use an artefact. An innocuous artefact, judged by its functional requirements, can be designed with a bad use of it in mind. This is no ground for calling the artefact bad, that is, something that should not have been designed, but still both the act of designing it for a bad purpose and the act of actually using it for that purpose would be wrong. Exactly these judgements seem defensible with respect to the atomic bomb had the Nazis developed and used one. And finally, an artefact may always be used in a way not intended or foreseen during its design. There is no contradiction in judging that the development of the atomic bomb by the USA and UK was not wrong, and that the resulting atomic bomb was not a bad artefact per se – two separate judgements – but that the use of two bombs on 6 August 1945 and 9 August 1945 was utterly wrong – a third judgement. This may well have been the opinion of many of the scientist and engineers who supported its development and contributed to its creation, since they considered it a weapon that would be used against the uniquely 'demonic' regime of Nazi-Germany, or even to keep this regime from using its own variant of it.

Conversely, there is also no contradiction in claiming, on the one hand, that certain artefacts should not be designed and brought into existence and in holding, on the other hand, that to use such an artefact can sometimes be right. In the

[9] Could it ever be different? Suppose that someone builds a copy of Kafka's torturing machine in order to see whether such a device was possible at all, or to see how it could work, or merely as an idiosyncratic way of expressing admiration for the story. Could this person claim that this copy was not a bad machine because it would not have been 'informed' by bad intentions? Surely we would judge that there was something definitely perverse about this motivation.

previous section it was suggested that even for bad artefacts like nasty weapons and instruments of torture, a positive form of use might be found. Here, however, the converse claim is that it can be right to occasionally use an instrument of torture for the purpose for which it was designed, to administer torture. To spell out under what conditions this could be right is not what is at issue here: my point is that one could occasionally feel justified in using an available instrument of torture for torturing, if only the likely benefits are good enough, and at the same time to hold that the instrument used should not have been available, because should not ever have been designed. Design is a action that results in an instrument *type*. The designer in principle makes available an unlimited number of copies, since any number of them could be produced either by following the blueprint or by copying them from a specimen produced. However much the designer may hope that the instrument under design will only be used in circumstances where its use will be justified, he or she has no control over its actual use, and by designing it makes the instrument available for as many morally unjustifiable uses of it as history may care to generate. This argument, to my mind, runs closely parallel to an argument presented by Seamus Miller concerning the question whether torture can ever be right. Miller argues (2011) that circumstances are conceivable in which torture may be justified – no a priori argument is possible, claims Miller, to show that it cannot possibly ever be right – but that still any government or state should make the application of torture by any of its citizens unlawful. To make legal room for torture, however finely the circumstances in which the law condones it are described, is to legalize a practice type, which is bound to exceed the cases in which a single act of torture is justified.

To identify bad artefacts as artefacts that it is wrong to design seems to me, then, the position that best matches the adopted view on the character of normative judgements and the relation between deontic and evaluative judgements. Wrong actions are actions whose resulting states-of-affairs are bad. Where judgements concerning the wrongness or rightness of actions are concerned, states-of-affairs are precisely the things one brings about. Shifting the focus to objects, artefacts are precisely the objects people bring into existence. One could even claim that the proposal defended here is a special case of the standard analysis, which links deontic judgements of actions to evaluative judgements on resulting states-of-affairs: the badness of a state-of-affairs consisting of a particular artefact having been brought into existence is transferred to the artefact whose new existence is seen as the crucial feature of the state-of-affairs brought about. However, to define a bad artefact as an artefact that one has a compelling reason not to design may be too simple, for the following reasons. Suppose one defines a design task by listing the functional requirements of a *perpetuum mobile*. One has compelling reasons not to undertake that task. Still, a *perpetuum mobile* is unlikely to be anyone's candidate for a bad artefact. A way out could be to question whether a *perpetuum mobile* is a kind of artefact in the first place, since it is an impossible object. Or, to take another possible counterexample to the analysis, suppose one is presented with a blueprint for a particular type of a perfectly normal artefact, say a washing machine, and suppose that the blueprint contains a serious flaw, which will result in a malfunctioning device. One has a

compelling reason not to build that thing. But again, which artefact, of what kind, has one a compelling reason not to bring into existence? If one accepts that this is a particular kind of washing machine, although one that is malfunctioning as a type, then we have a case of an artefact that one should not bring into existence but that is, arguably, not a bad artefact. Clearly there are subtleties that have to be sorted out on the way to a precise definition in terms of necessary and sufficient conditions, but this is not the place to further discuss these subtleties.

In the next section I investigate whether the analysis adopted here carries over without change to attribution of non-instrumental *good*ness to technical artefacts. I will argue that the situation is to a large extent but not entirely symmetrical, which may explain why examples of quintessentially bad artefacts are more salient than examples of quintessentially good artefacts. In the final section I investigate whether the analysis adopted here could also be applied to other artefacts than technical artefacts, in particular to works of art. I develop a proposal on how it could, which builds, inevitably, on a specific view of the character of artworks. The next two sections, then, are of a more explorative sort than the previous ones.

## 12.5 The (A)symmetry of Goodness and Badness

In the introductory section, I stated that quintessential examples of bad artefacts easily spring to mind but that examples of quintessentially good artefacts do not pose themselves equally easily. Indeed, von Wright's criterion of bad artefacts being artefacts that have detrimental side-effects does not straightforwardly leave room for the definition of good artefacts. Are these to be artefacts that have beneficial *side* effects? Even if their *effects* are typically bad? Both the differences in the psychological salience of bad vs. good artefacts as von Wright's proposed criterion for what makes artefacts bad suggests an asymmetry between good and bad artefacts. It seems difficult, however, to square this with the proposed analysis of the (moral, non-instrumental) value objects can be said to have.

On this analysis, if bad artefacts are artefacts one has a conclusive reason not to bring into existence, and perhaps, once they exist, to avoid, destroy, and so forth, then good artefacts are artefacts one has a conclusive reason to bring into existence and once they exist, to admire, cherish, protect, and the like. As a consequence, it seems, we must admit that most artefacts are good. Artefacts generally make our lives go better: that is why we have them and use them. Even if many artefacts perhaps only make our lives go marginally better, still if there were no other reasons against bringing them into existence, then even making our lives go marginally better would suffice for tipping the balance toward creating them.[10] This, however, introduces conceptual difficulties. A bad artefact is a device that should not be designed;

[10] There is of course the question whether it is true that artefacts make our lives go better. This is an aspect of the issue whether there is such a thing as 'progress'. I cannot take up this question here. For an interesting treatment, see Rescher (1980).

specifying whether the option of designing it is available does not change this. This is hardly true for good artefacts: these are not devices that should be designed, period. We do not hold it against James Watt, or anyone prior to Faraday, that they did not design the dynamo. So by 'should be designed', I will mean that if the option of designing a specific artefact, individuated minimally by functional requirements as sketched in the previous section, presents itself, there are strong reasons for choosing that option.[11]

Merely contributing positively to the ways our lives go, or being supported by the balance of reasons tipping towards creating it, may seem a meagre basis for calling an artefact good, however. As I construe it, being bad for an artefact is like being tainted by human badness. One would, then, require the goodness of an artefact to consist at least partly in its being touched by human goodness. But many artefacts that make our lives go better have been created without a specific intention to create an artefact that would make lives go better. No specific effort to establish this was part of the development process; instead, they were made purely because a market for them was perceived to exist. We could decide, therefore, not to enforce a complete symmetry between the goodness and badness of artefacts. A bad artefact is an artefact that one should not design. The reasons are primarily grounded in the badness informing the functional requirements; it would be tainted in this way were it to exist. A good artefact is an artefact that one should design. But one need not claim that this is necessarily so because of the goodness that informs the functional requirements; the good it could do if it were to exist might be considered to serve as a reason just as well. It is conceivable that the functional requirements do not reflect at all this good it could do, nor the intention that it do this good. In principle it could even occur that, for the way it could make our lives go better, one judges of an existing artefact that it should have been designed, although it was actually designed for the purpose of making at least some lives go worse, although I have no plausible example of such a case. Then the symmetric part of the analysis would be that an artefact is bad if one should not bring it into existence and an artefact is good if one should bring it into existence. The asymmetric part would be that the reasons for which an artefact should not be brought into existence may be grounded in different kinds of properties or features compared to the reasons for which an artefact should be brought into existence. For a bad artefact, bad intentions informing its design are required, but for a good artefact explicitly good intentions informing its design would not be required; in the absence of such intentions – and of course in the simultaneous absence of bad intentions – good forms of use could be sufficient.

One should not conclude that on the proposed analysis any artefact is either good or bad. It would be rash, for instance, to conclude that the goodness of artefacts, as

[11] The reasons, however, cannot be compelling, because for instance one of the other options on the table may be to design a thing that is superior to the one under consideration. How to conceive of this modal notion of 'should be designed' is highly dependent on how we individuate the artefact kind at issue and how we individuate the act of designing it. This is an issue where there is still much unclarity, and I cannot satisfactorily discuss it, let alone solve it, in this paper. The issue returns in the next section.

objects designed to make our lives go better, extends to such artefacts as weapons. The arguments presented in the previous section for the ultimate relevance of the precise formulation of the design requirements and their motivation may do work here as well. A bomb, even an atomic bomb, may be used to good purposes, for example by keeping a rogue terrorist nation in check. But an atomic bomb is a bomb, not a keeping-rogue-nations-in-check-er. It has been designed as an instrument to cause gigantic explosions, resulting in destruction on a massive scale. That sort of instrument is not one that can be expected to make our lives go better. Nuclear weapons are at best necessary evils. The verdict is less clear for ordinary weapons like guns. Presumably, by allowing the police to keep criminals in check, they do make our lives go better. Guns can arguably be seen as keeping-bad-people-in-check-ers. But we only have a need for such things because bad people use guns as making-good-people-serve-my-wishes-ers. It is a bit nonsensical to see guns as the former and consider the latter their accidental use. Here the often disputed neutrality of artefacts really has a bite. Because guns are used widely for both good and bad purposes, and their design puts no obstacle to either form of use, nor suggests anything concerning which of the two is 'proper' use, we have no basis for saying either that guns are bad or that they are good.

The relation between an artefact's goodness or badness and the reasons one may have for or against using it seem to show no asymmetry with respect to good or bad. If an artefact is bad, one has a reason not to use it for its purpose, but not more reason than one has reason not to use anything for that purpose. For torturing someone, one could just as well use the contents of an ordinary toolbox, as happens, for instance, in Guillermo del Toro's film *El labirinto del fauno* (2006). These tools are not bad, but one has neither more nor less reason to use them for torturing as one has reason to use a device designed for torturing, which to me is a bad device. The reasons grounding the badness of a torturing device are reasons for its (would-be) designers. Likewise, if an artefact is morally good, one has a reason to use it when the corresponding need occurs, because the way its presence makes our lives go better is through the possibility we have of using it,[12] but a (prospective) user does not have more reason to use it than to use anything that would serve equally well, instrumentally speaking. This remains true even if restricted to proper use. A particular public waste receptacle may be termed morally good, minimally in a comparative sense morally better than other waste receptacles, if due to some of its features it invites people more often to actually dispose of their waste in it than other receptacles do. In this way the receptacle makes the lives of everyone go even better (if perhaps only slightly) than other public waste receptacles do. However, a passer-by who is looking for a place to dispose of the remains of a take-away lunch does not have more reason, or even a reason, to throw them into the inviting one rather

[12] There will be some fine-tuning at work in the background: the need must itself be reasonable, otherwise one cannot have a reason to satisfy it, and the particular artefact must be operational and not malfunctioning, otherwise one cannot have a reason to use this artefact. See Franssen (2006) and (2009) for more details concerning this fine-tuning.

than any other one.[13] The receptacle's features are reason-giving for designers and for implementers, i.e., agents who make them available for use, e.g. municipalities, but not for users.

## 12.6 Technical Artefacts vs. Artworks

Up till now I have dealt only with the moral value of technical artefacts. However, technical artefact are not the only objects for the existence of which we are responsible and which are eligible for normative judgements. Such judgements are applied to artworks as commonly as they are to technical artefacts: we have good paintings and poor paintings, beautiful sculptures and ugly sculptures. Naturally one may wonder whether the approach sketched in this paper to account for both judgements of moral and of instrumental value can be extended to works of art. I will briefly argue that it can.

Although space does not allow me to work out the details here, my position is that the typical value judgements applied to artworks quoted above behave like the corresponding evaluative judgements for technical artefacts and accordingly that aesthetic value is quite similar to the instrumental value of technical artefacts. Works of art are created by their authors in order to be used by their 'users', i.e. their audiences – here to be interpreted, etymologically incorrectly, as consisting of listeners and/or viewers – to generate certain psychological responses, partly emotional and partly cognitive responses. A prime task for such a view is to characterize the nature of these responses; they have to be distinguished on the one hand from the enjoyment provided by sexual gratification or by the consumption of ordinary food when hungry – but I do allow for *haute cuisine* due to the sort of response it aims for to fall under art – and on the other hand from responses like those caused by a mock execution or by scaring the living daylights out of someone. Here, however, I will just assume that such a characterization is possible. What is more, with the developments of arts during the past century, artists have also themselves become the users of their works by confronting people who have not volunteered to enjoy them in order to provoke an emotional or cognitive response. Whereas instruments are meant to be wielded by users to transform *matter*, art works are wielded to transform the *soul*, to use a bit of old-fashioned language that seems apt here.[14] This view furnishes an explanation of why 'beautiful' and 'good' are very close to each other in the positive evaluation of a work of art and 'ugly' and 'poor' in the

[13] There is a whiff of paradox here. How can the receptacle be inviting if one does not have a reason to use it? But there is actually no conflict, because 'inviting' should be interpreted as a behavioural term, which does not require reasons or intentionality.

[14] Stalin famously addressed writers as 'engineers of the soul' during a meeting in 1932, to express his wish for an 'industrialization' of literature in support of the building of socialism. He may have borrowed the expression from the Russian avantgardist author Sergei Tretiakov, who in 1926 had referred to writers as 'psycho-engineers'. See Golomstock (2011, pp. 26 and 84).

negative evaluation, and even why 'good' and 'poor' have tended to replace 'beautiful' and 'ugly' during the development of modern art, which is characterized partly by a considerable widening of the sorts of responses that a work of art can aim for. On this view, a good artwork is one likely to be successful in provoking a particular response within the spectrum aimed for, and a poor one is one unlikely to provoke such a response, typically because it is likely to provoke an unintended, conflicting response. Accordingly, on my analysis of normative statements as being judgements expressing second-order facts, a good work of art is a work of art that someone who is after a particular emotional or cognitive experience has a reason to 'consume', and a poor work of art is one that someone who is after such a response has a reason to avoid consuming.[15]

Do works of art behave similar to technical artefacts also with respect to non-instrumental moral value? In popular writing there is no shortage of candidates for bad artworks. Pornographic books and films, books, films and computer games extolling (extreme, arbitrary) violence, Leni Riefenstahl's film *Triumph des Willens*, to name just a few.[16] However, to call these bad works of art is to judge them for their side-effects. The considerations that I presented above to argue that bad side-effects form insufficient ground for terming artefacts bad apply here as well, or even with more force. The bad effects of any of these candidates are contested, to say the least, and if they occur are highly dependent on the historical and cultural settings. One can guard oneself against suffering the undesirable side-effects, or be educated into (relative) immunity, consume antidotes, and so forth, strategies that usually work less well or not at all against the bad side-effects of technical artefacts. Riefenstahl's *Triumph des Willens*, being obviously technically antiquated and 'historical', reminds of the horrors of Nazism, and perhaps testifies to moral flaws in its author, but is no longer considered capable of corrupting its viewers; instead the film is now often admired for its instrumental excellence.

Are there, then, examples of works of art for which we rightfully say they should not have been created, for the same reasons that technical artefacts like instruments of torture should not be created, because they were specifically and exclusively intended to harm people? To physically harm its consumers seems a self-defeating strategy for a work of art; more plausibly a work of art harms by poisoning the minds of its consumers against other people. Naturally Hitler's *Mein Kampf* comes to mind. If this example is controversial, it probably is because it raises the definition of what we mean by a work of art. However, if we exclude it on grounds of missing aesthetic appeal and pretence, then we will end up with, minimally, a third subclass within the category of artefactual objects, namely the class consisting of non-fiction artworks: argumentative books, documentary film, spoken argument, and the like. For the sake of the present argument, I will cut this discussion short by treating these as artworks. Then, if one accepts *Mein Kampf* as an artwork analogue to Kafka's torturing machine, as a book that should not have been written, should

[15] I would even claim that for a relatively narrow set of traditional responses associated with artworks, 'beautiful' and 'ugly' convey the same normative content as 'good' and 'poor' do.

[16] Some people might even want to include Kafka's 'In the penal colony'.

not *Triumph des Willens* not equally be condemned as a film that should not have been made *in 1935*, because it supported the world view of *Mein Kampf*? This, I think, would be conflating the judgements that Leni Riefenstahl's *act* of directing the film in 1935 was *wrong* and the judgement that the film is *bad* as an object, a work of art. I see no compelling reason for the latter judgement, since the film itself does not partake in the Nazi business of harming, nor, as far as we know, did Riefenstahl support that aspect of Nazism.[17] It is, in my view, clearly different from an instrument of torture, which keeps breathing an eagerness to inflict pain and a capability of doing so, even if we place it inside a museum.

To distinguish between the act of designing an artefact and the artefact itself may be considered more problematic for artworks than for technical artefacts. Can the option of making *Triumph des Willens* be considered for anyone but Leni Riefenstahl in 1934–1935? I take it, however, that artworks can be individuated through a combination of intentions of a specific sort and a specific material realization, and do not emerge from specific historical acts as a matter of metaphysical necessity.[18] If there is a difference with technical artefacts, it is that their material realization contributes to their individuation at a level much more detailed than is the case for technical artefacts. In fact, artworks are generally considered to be individuated as unique objects, not as representatives of types.

The differences between technical artefacts and artworks are especially relevant to positive judgements of artefacts. Like many technical artefacts, many artworks enrich our lives, which makes them non-instrumentally good. Consequently there are strong reasons to design or create a specific artefact if that option is on the table, and strong reasons for adopting specific positive attitudes with respect to existing ones. For artworks, these positive attitudes or actions include caring for them, protecting them, and similar ones, more forcefully than for technical artefacts. If we judge that destroying a particular technical artefact is morally wrong, it is typically because it is owned and we in fact judge the act of harming its owner. If this aspect is removed, and a technical artefact is destroyed by its owner, or with the consent of its owner, this will usually not be felt as morally wrong. But when the Japanese businessman Ryoei Saito, who had bought Van Gogh's painting 'Portrait of Dr. Gachet' in 1990 for the then record price of 82.5 million USD, suggested that he would have the painting cremated jointly with his body after his death, this caused an outcry, and I suppose there is wide agreement that it would have been wrong of him to do so.[19] To destroy an artwork is usually to deprive all of humanity of the possibility of using and enjoying it, whereas technical artefacts exist often in so

[17] One might claim that, by celebrating the Nazi image of the blond and tall Arian, the film partook in the Nazi definition of the *Untermensch*. Then why not label the average Hollywood film or U.S. television series as morally bad for celebrating an ideal of human beauty that the majority of the population falls grossly short of? Because we do not kill plain people? Neither was there any systematic killing of *Untermenschen* in 1935.

[18] Cf. the story by Jorge Luis Borges, 'Pierre Menard, author of the *Quixote*'.

[19] Eventually it was not cremated with him, although for almost a decade after Saito's death the fate of the painting remained a mystery.

many copies that the loss of one hardly affects the availability of this particular kind, or, in case where only a few copies exist, a new, instrumentally equivalent copy can in principle always be made. Nevertheless, it may be questioned whether their uniqueness is an essential feature of artworks. Prints and recordings of musical works are not and are none the less artworks for that. It is typically in the interest of the authors to secure the uniqueness of work of art, primarily because it is in the interest of their customers, but still many painters have made copies of their own work, occasionally on request.

To insist that it is the author's prerogative to decide about the number of copies that exist of a work of art is a compromise between the judgements that the availability of an artwork, as an enrichment of our lives, is a good thing and that a specific regime for allowing the author of an artwork to reap the (monetary, reputational) benefits of authorship is also a good thing. Which still leaves considerable room for deciding what the precise form of that compromise should be, and the precise regime for an author's possibilities at benefitting. However that issue is decided, it seems then that if artworks do indeed enrich our lives, the availability of (photographic) reproductions, which do not infringe on the intentio-causal relation between the original work and its author, is also good, since they share to some extent, depending on the quality of the reproduction, the capacity to enrich our lives that the original work has, and this irrespective of whether the author of the original cares to agree. Continuing this line of reasoning, it would be good, and there would be strong reasons in favour of realizing it, that near-perfect copies existed of major works of art, which could replace them if they were destroyed. This is already common practice for outdoor statues and sculptures, but there are good reasons for practising this more widely (which is not to say that there are no reasons against it). The reconstruction of historical buildings destroyed by fire or war or natural disasters like earthquakes is similarly motivated. Such copying and reconstruction is less controversial the less intimately the hand of an author has touched the work and is still felt to rest there. It is a serious question how much moral weight this should bear in deciding on how to secure the availability of artworks; perhaps it should bear weight only to the extent that this intimacy determines their value to us, their 'users'.

## 12.7 Conclusions

Let me finally merely briefly restate the gist of my claims. Our common talk of good and poor artefacts concerns their instrumental value, their goodness or poorness in realizing the purpose for which they were designed. Instrumental goodness or poorness is different from non-instrumental, moral goodness or badness. With respect to non-instrumental value, we can term some of the artefacts in existence bad and many of them good. This is not because of a surplus of goodness in the world, but because we are ourselves responsible for bringing artefacts in existence and we are sufficiently rational for bringing them in existence roughly in accord with the

reasons we have for doing so. Further, even though most artefacts are good, using them can, in particular circumstances, be wrong. Conversely, even though some artefacts are bad, using them can, in particular circumstances, be right.

## References

Dancy, J. (2000). Should we pass the buck? In A. O'Hear (Ed.), *The good, the true and the beautiful* (pp. 159–173). Cambridge: Cambridge University Press.

Dancy, J. (2006). Non-naturalism. In D. Copp (Ed.), *The Oxford handbook of ethical theory* (pp. 122–145). Oxford/New York: Oxford University Press.

Franssen, M. (2006). The normativity of artefacts. *Studies in History and Philosophy of Science, 37*, 42–57.

Franssen, M. (2009). Artefacts and normativity. In A. Meijers (Ed.), *Philosophy of technology and engineering sciences* (Handbook of the philosophy of science, Vol. 9, pp. 923–952). Amsterdam etc.: North-Holland.

Franssen, M. (2013). The goodness and kindhood of artefacts. In M. J. de Vries, S. O. Hansson, & A. W. M. Meijers (Eds.), *Norms in technology* (Philosophy of engineering and technology, Vol. 9, pp. 155–169). Dordrecht/Heidelberg: Springer.

Golomstock, I. (2011). *Totalitarian art in the Soviet Union, the Third Reich, Fascist Italy and the People's Republic of China* (R. Chandler, Trans.). New York/London: Overlook Duckworth.

Hansson, S. (2006). Category-specified value statements. *Synthese, 148*, 425–432.

Miller, S. (2011). Torture. In E. Zalta (Ed.), *Stanford encyclopedia of philosophy*. http://plato.stanford.edu/entries/torture/

Rescher, N. (1980). *Unpopular essays on technical progress*. Pittsburgh: University of Pittsburgh Press.

Scanlon, T. M. (1998). *What we owe to each other*. Cambridge, MA: The Belknap Press of Harvard University Press.

Thomasson, A. H. (2003). Realism and human kinds. *Philosophy and Phenomenological Research, 67*, 580–609.

von Wright, G. H. (1963). *The varieties of goodness*. London/New York: Routledge & Kegan Paul/The Humanities Press.

Winner, L. (1986). *The whale and the reactor*. Chicago: University of Chicago Press.

# Chapter 13
# Values in Chemistry and Engineering

**Sven Ove Hansson**

**Abstract** There are substantial similarities in how value statements are applied to chemistry and technology. Both disciplines are subject to negative moral valuations due to the harmful effects of some of their products. In addition, instrumental value statements of a specific type, namely category-specified value statements, are used in both areas. Examples are "a bad engine" and "a good stabilizer". In both cases this usage is based on functional descriptions that relate to the design component of the respective discipline. However, there are also important differences in how such value statements are applied in chemistry and in technology. The similarities and differences are investigated, and it is concluded that additional studies along these lines can contribute to our understanding of both disciplines.

## 13.1 Introduction

Engineering design and chemical synthesis are closely related forms of human agency: they both aim at constructing objects that do not exist in nature. In other words, they are two branches of the sciences of the artificial. After a discussion of the nature of this common feature in Sect. 13.2, the rest of this article will be devoted to adding a further aspect to the comparison between chemistry and technology: that of values. It will be shown that the element of design in both these human endeavours has led to important similarities in the value statements that are made in relation to them. These similarities are of two major types. First, and perhaps most obviously, design activities are subject to *moral* evaluation of a kind that the search for knowledge per se is usually not exposed to. Therefore, chemistry and technology

S.O. Hansson (✉)
Department of Philosophy, Royal Institute of Technology (KTH), Stockholm, Sweden
e-mail: soh@kth.se

P. Kroes and P.-P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Philosophy of Engineering and Technology 17, DOI 10.1007/978-94-007-7914-3_13,

are frequent targets of such appraisals. This will be the subject of Sect. 13.3. Secondly, there is a discourse on the *instrumental* values of our constructions, such as when we talk about a good hammer or a bad catalyst. This will be the subject of Sect. 13.4. Section 13.5 is devoted to a brief discussion of artefacts as assigners of moral value. Some more general conclusions are offered in Sect. 13.6.

## 13.2 Two Sciences of the Artificial

Technology has sometimes been described as applied natural science (Bunge 1966), but today the general consensus, at least among philosophers of technology, seems to be that the technological sciences are different in several respects from the natural sciences (Mitcham and Schatzberg 2009; Bunge 1988; Hansson 2007). One essential difference, perhaps the most basic one, concerns its study objects. The study objects of the natural sciences are objects and processes in nature, such as atoms, animals, volcanoes, chemical reactions and the evolution of species. The study objects of the technological sciences are objects and processes that are deliberately constructed by humans, such as combustion engines, airplanes, computer programs, light bulbs, landmines, and artificial heart valves. The main purpose of projects in the technological sciences is often to construct or modify such objects, rather than to study pre-existing ones.

The distinction between natural and man-made objects of study is simple, and it is useful for distinguishing between the natural and technological sciences. But, concededly, it is somewhat oversimplified and does not fit perfectly with actual linguistic practices. At least three general caveats have to be added.

First, this definition refers to the ultimate study objects, not necessarily to the actual physical objects that are subject to observation and experimentation. Natural scientists often study objects that have been modified for the purpose of measurement or experiment, such as a crystallized version of a protein that does not appear in crystallized form in nature. However, this is done in order to better understand the composition and the properties of naturally occurring objects (in this case the structure of the naturally occurring, uncrystallized protein). In the technological sciences, the humanly constructed artefact, such as a machine part or a computer program, is the ultimate study object (Hansson 2007).

Secondly, only some human constructions belong to the domains of the technological sciences. Others belong to the social sciences, the humanities, or mathematics. As one example of this, money is a human construction, but most aspects of this construction belong to economics rather than to the technological sciences. The central concern of the technological sciences is the construction and the immediate usage of material objects (and of immaterial objects such as computer programs that are needed to employ some of these material objects in the intended ways).

Thirdly, it has to be recognized that the same physical laws apply to both natural and artificial objects. This is, of course, the reason why we can use artificially constructed objects in experiments that aim at uncovering the workings of nature.

It is also the reason why we can apply laws of nature in the study and construction of technological objects, e.g. thermodynamic theory in the construction of combustion engines and silicon chemistry in the construction of solar cells. However, even though (applied) natural science is useful – and indeed indispensable – in technology, it is not sufficient. Direct studies of the technological artefacts themselves are also necessary. The reason for this is, fundamentally, that science is a human enterprise, and it must therefore be conducted in terms that are cognitively accessible to humans. An example: We have good reasons to believe that the human heart operates in full accordance with the fundamental laws of physics, but there is no way to describe the workings of the heart directly in terms of physical laws – at least not if the description is required to be useful for human understanding. In order to understand how the heart works we need to analyze it in the terms used in physiology, such as atrium, sinoatrial node, and systole that are not definable in the terms of physics. Similarly, a humanly understandable account of a computer will have to use intermediate-level terms such as register, central processing unit, and memory. As the latter example shows, the distinction between the technological and the natural sciences depends on the limitations of human cognition (the former example shows that the same applies to the distinction between the different natural sciences).

In addition to these three general caveats a special exception seems to be needed for one of the major branches of natural science, namely chemistry. As was pointed out by Schummer (1997), chemistry is in some respects closer to technology than the other natural sciences. The crucial similarity is the role of design in these two areas. Just like engineers, many chemists construct their own (ultimate) objects of study. Chemical synthesis is analogous to engineering design, and the study of synthesized substances is analogous to the study of (designed) technological objects. Chemical synthesis is usually performed for instrumental reasons, just like engineering design, but in some cases it takes place as an end in itself (Schummer 2001). We can describe technological science and (large parts of) chemistry as two branches of the sciences of the artificial. (Chemistry and engineering are combined in interesting ways in chemical engineering, a discipline that will not, however, be treated here.)

## 13.3 Moral Valuations

One important similarity between chemistry and engineering is that they both have serious public relations problems. In the eyes of large segments of the public, both chemistry and engineering are associated with environmental problems and other failures of science and technology (Heilbronner and Wyss 1983; Berdonosov et al. 1999; Becker 2009). In most of the Western world, recruitment to the engineering profession is weak. The number of young people who choose the profession of an engineer is too small to cover society's need for engineers. This problem has often been attributed to a negative attitude among the public to technology and engineering (Anderson-Rowland 1996; Yurtseven 2002). The public seems to associate

technology with its failures, perhaps in particular its harmful environmental consequences, rather than with its positive achievements. The very word "engineering" seems to have acquired a negative connotation, as can be seen from how the phrase "social engineering" is used as a derogatory term about criticized social practices (Hansson 2006c). Understandably, young persons choosing a profession look for a vocation that is well respected and generally conceived as valuable for society. In many of the technologically advanced countries of the world, engineering is not such a profession.

Chemistry is very much in the same position as engineering with respect to its public relations. At least since *Silent Spring*, chemistry is largely perceived as a pernicious activity. The public seems to associate chemistry with synthetic products that poison the environment and pose threats to human health. Marketing practices are a sure sign of this. Consumer products have been marketed with the (scientifically inaccurate) phrase that they contain "no chemicals". Educations in chemistry have the same type of recruitment problems as engineering educations (Read 2010).

It is interesting in this context to note a shared feature of chemistry and engineering: they both involve the construction of new kinds of material objects that did not exist before, neither in nature nor as artefacts. With a positive phrase such activities can be called "innovative design". However, when they are perceived as problematic they are more often called "unnatural" or "tampering with nature" (Sjöberg 2002).

One might think of "natural" and "unnatural" as descriptive terms, but they are in fact value-laden to a very high degree. If you say that the aggressive behaviour of a child is natural, then that will be understood as an expression of acceptance. If, on the other hand, you say that a food additive is unnatural, then that will be understood as an indication that it is an undesired ingredient. Indeed, the term "unnatural" seems almost never to be used about phenomena that are generally accepted. I never heard anyone call it unnatural to wear eyeglasses or to boil contaminated water before drinking it. In contrast, pasteurization has been called unnatural and so, in certain religious circles, has the use of condoms.

Hence, to say that something is "unnatural" is not a mere a statement of fact but also a way to express a negative valuation. Presumably, most of those who condemn homosexuality as "unnatural" would not change their view if it could be proven to them that human beings have a biologically based tendency to homosexuality.

There seems to be, in public perception, a particular form of badness that consists in transgressing the perceived boundaries of what nature allows. In part, this may be seen as a secularized idea of religious origin. In many religious traditions, God-given laws are said to be written into nature. According to Tertullian "quod Deus noluit utique non licet fingi" ("what God did not want definitely cannot be allowed to be invented"). Objects not found in nature "a Deo non sunt, auctore naturae. Sic a diabolo esse intelleguntur, ab interpolatore naturae" ("do not come from God, the Author of nature. They must therefore be understood as coming from the Devil, the disrupter of nature") (Tertullianus ca. 200, Book I, Ch. VIII.).

As an example of this, said Tertullian, God has not created purple sheep. Therefore he did not intend us to dye our clothes to make them purple. On the other hand, Tertullian did not oppose the use of ploughs, books, or looms, none of which

can be found in nature. Subsequent critics of "unnatural" technology and chemistry have been similarly selective. However, both technology and chemistry are so often accused of being "unnatural" that this criticism has had an effect on the general perception of these disciplines. As was noted by Schummer, a dichotomy between "chemical" and "natural" is usually taken for given, but nobody would suggest a dichotomy between "physical" and "natural" or between "biological" and "natural" (Schummer 2003).

But contrary to Tertullian, modern critics of "unnatural" technology and chemistry often have a point that needs to be taken seriously. The technologies and chemical practices most often criticized for being "unnatural", such as biotechnology, nanotechnology, and the use of pesticides, all have in common that they are associated with uncertainty. The introduction of such new technologies often involves the use of new constructs whose potential effects on human health and the environment cannot be accurately predicted from previous experiences. This is a reason to take special care in the investigation of their possible adverse effects, and in many cases also to take precautionary action. However "unnaturalness" is a much less useful specification of the problem than "uncertainty", not least since the latter is more readily accessible to rational argumentation (Hansson 2003).

## 13.4 Instrumental Valuations

In order to understand the specific nature of some of the value statements made in technology and chemistry we first need to consider a prominent feature of the language we use when discussing technology and chemistry, namely the use of functional descriptions. Since both technology and (large parts of) chemistry are concerned with designing objects to be used in predetermined ways, functional descriptions have a prominent role in both these disciplines.

### *13.4.1 Functional Descriptions*

Technological objects can be described either in terms of their physical–structural characteristics or in terms of their functions ("wooden cylinder, 3 mm thick and 94 mm in diameter" – "cup coaster"). According to the "dual nature" theory of technological artefacts, they can be understood both as physical objects and as objects with certain functions. Whereas the physical properties of a technological object can be described without any reference to human intentions, its functional properties are closely related to the intentionality of design processes (Kroes and Meijers 2006; Kroes 2006; Vermaas and Houkes 2006; Hansson 2006b).

Some of the categories that we use to categorize technological objects are purely functional. Nutcrackers, calculators, pens, airplanes, and CPUs are examples of this. A device with the function to crack a nut can be called a nutcracker, irrespectively

of its physical structure. The defining characteristic of a purely functional category is that in order to determine whether an object belongs to that category it is sufficient to ascertain its function. Hence, in order to determine whether an object is a plough we have to find out whether or not its function is to turn over the upper layer of the soil. In order to determine whether a computer program is a search machine we have to find out if it serves to find digitally stored information. We do not need to find out what its components are or how they have been combined.

Other technological categories are predominantly structural. This applies for instance to the notions of a plank, a steel wire, a rope, and a fibreboard. As these examples indicate, technological categories defined in structural terms tend to be raw materials or multi-purpose components. The defining characteristic of a structural category is that in order to determine if an object belongs to it, it is sufficient to know its structure, i.e. what its components are and how they are put together (however, as was pointed out to me by Peter Kroes, when we describe an object in purely structural or physical terms we are, strictly speaking, referring to it as a physical rather than a technological object).

There is also a third type of categories of technological objects, namely those whose definition combines functional and structural characteristics. The notion of "scissors" is an example of this. Saws, knives, and scissors have very similar functions; to distinguish them we need to refer to their structural properties. We call a cutting instrument a pair of scissors only if it has two edges that can slide against each other; furthermore its cutting function must rely on that sliding of edges. Cogwheels are another category of the mixed type. We would not use that term for a toothed wheel that was constructed for some other purpose than to connect it with another toothed device so that movement of one of them induces movement of the other. Neither would we use it for an untoothed wheel that connects with another wheel through some other mechanism. The defining characteristic of this mixed type of categories of technological objects is that in order to determine whether an object belongs to such a category it is necessary to have information about both its structure and its function.

The categorization of technological objects can often be performed in several steps, in categories and subcategories. In the creation of subcategories, structural or functional categories are often subdivided into subcategories of the mixed type. Hence "engine" is a functional category but "two stroke engine" a mixed one. "Clock" is a functional category but "pendulum clock" a mixed one. "Plank" can be defined as a structural category ("a piece of sawn timber at least 50 mm thick and 225 mm wide" according to the Oxford English Dictionary), but "floor plank" is certainly a mixed one. Some categories can be divided into subcategories both in terms of structure and function ("pipe" – "copper pipe" – "sewage pipe"). In summary, engineers operate with terminologies that are based on complex mixtures of functional and structural specifications. In this respect, the "dual natures" of technological objects are intertwined rather than juxtaposed.

Just like technology, chemistry operates with both functional and structural categories. Obviously, the most fundamental chemical categories, those representing particular chemical substances, are structural. We define chemical substances

according to the structure of their molecules. However there are at least two important additional categories of chemical terminology that are often functional, namely those referring to groups of substances and to parts of molecules.

Beginning with groups or categories of substances, chemists operate with both structural and functional categories. "Sulphur compound", "organic compound", "oxide", "salt", "crystal", and "polymer" are all defined in terms of molecular structure or (in the case of "crystal") spatial configuration. But there are also many functional categories of substances, such as "solvent", "catalyst", and "oxidizer". The use of such terms is closely related to the laboratory activities of a chemist. A solvent is a substance you can use in the laboratory to form a homogenous mixture with other substances, a catalyst is a substance that can be used to change a reaction rate without itself being consumed, etc.

Organic chemistry has a well-developed terminology for parts of molecules, usually substituent groups but also for instance "benzene ring" and "haem *b* group". When a part of a molecule is completely specified, this is done in structural terms: butyl group, formyl group, carbonyl group, phenyl group etc. Classes of such structures are also usually classified in structural terms (alkyl groups, halide groups). However, although these structures and classes of structures are defined in structural terms, the classifications in question are commonly called "functional groups" (or "moieties"). A functional group is "an atom or group of atoms which has a characteristic effect on the physical or chemical properties of the molecule to which it belongs" (OED). The reason why it is called "functional" is of course that through its inclusion a molecule acquires properties such as being able to react with certain other molecules; these properties can be called "functions". Functional groups are probably best described as definable in terms of a combination of functional and structural characterizations. Ideally, the functional and structural characterizations used in organic chemistry are meant to correspond to each other.

In conclusion, both chemistry and technology make ample use of both functional and structural terms, and they also both employ terms that combine the two forms of specification in intricate ways. The reason why functional terms are used is essentially the same in both areas: both engineers and chemists design or choose objects for a variety of purposes, and there is a need for terms that apply to the objects that can be used for particular purposes. However, there are also differences. Functional notions have a somewhat less prominent role in chemistry than in technology. Furthermore, a major group of functional terms in chemistry, namely those referring to functional groups (moieties), are closely correlated with structural properties. Such close correlations between function and structure are not easily found in technology.

### *13.4.2 Category-Specified Value Statements*

Value statements should be understood as relative to some more or less explicit criterion or standard of evaluation. The criteria commonly referred to provide us

with types of value statements such as instrumental values, aesthetic values, ethical values, etc. (von Wright 1963). Values that are specified in this way can be called *viewpoint-specified*. There are also value statements that are *unspecified* with respect to the criteria of evaluation since they are intended to represent an evaluation that takes everything into account. Such values have also been called "synoptic" (Rescher 1968, p. 293) or "categorical" (Rawling 1990, p. 495). As I have argued elsewhere (Hansson 2006a), there is in addition a third group, namely *category-specified* value statements. These are value statements that are specified in terms of some category that the object of value belongs to. Examples are easy to find in everyday conversations:

She is a good dancer.
The bicycle that I bought was a really bad one.

Category-specified value statements tell us what value something has as a member of a specific category. In the first of the above two examples, the value-specifying category is that of a dancer, in the second that of a bicycle. The value-specifying category is an essential component of the evaluation. To be a good dancer is not the same as to be good and also a dancer. Obviously, an object of value can belong to two different categories and be evaluated differently in them. A good dancer can be a bad singer, and the other way around.

A large part of the value statements that we make about technological artefacts are category-specified. We speak about "good" cars and "bad" computer programs, and engineers constructing new devices look for "better" components and materials. Category-specified value statements about technological objects can be divided into three major groups according to whether they refer to a purely functional category ("a good hammer"), a mixed category ("a worthless pair of scissors"), or a structural category ("an excellent steel wire").

When the value-specifying category is purely functional, then values are assigned according to how well that function is satisfied. Hence, a hammer is an object with the function of driving nails or striking blows at material objects. It follows from this that a good hammer is one that satisfies this function well, so that blows can be struck with maximal precision and minimal effort. In general, if we know the functional criterion in terms of which a category $X$ of technological objects is defined, then we also have the criteria for a good $X$ or a bad $X$. In these cases, being a good $X$ means to fulfil the defining functions of an $X$ to high degree, being a bad $X$ means to perform them to a low degree, and similarly for other value terms.

Next, let us consider cases when the value-specifying category belongs to the mixed type, such as pairs of scissors. As already mentioned, a pair of scissors is characterized by having (1) the structural characteristic of two edges that can move along each other, and (2) the function of cutting. When we talk about a good or bad pair of scissors, we refer to their capacity to cut, i.e. to their functional characteristic. More generally speaking, in the mixed cases the valuation refers to the functional component of the defining characteristic of the category.

However, in the mixed cases the structural component also has a role in the evaluation, namely in setting the standards for what is good or bad satisfaction of the criterion. Roughly, we can say that the functional component determines the

scale of measurement whereas the structural component determines the limits on that scale. As an example of this, a good plastic hammer is a hammer that, in comparison with other plastic hammers, satisfies the (general) functional requirements on a hammer to a high degree. The criterion is the same for a plastic hammer as for a steel hammer, but the level required for goodness is presumably lower for plastic hammers. (However, in some contexts a wider comparison class is referred to. The sentence "there are no good plastic hammers", means that there are no plastic hammers that, in comparison with other hammers in general, satisfy the functional requirements on a hammer to a high degree.)

Our third group of category-specified value statements are those that refer to some category of technological objects that are categorized in terms of their structural properties. Such valuations are very common and they are also easily understood. For instance, we have no difficulties in understanding what is meant by a "good" steel wire: it should answer to certain specifications that correspond to what is required in most applications in which a steel wire can expectedly be used. In this way, artefact-types that are defined according to their structure are nevertheless evaluated in terms of the functions for which they are typically used. Hence, all three types of category-specified value statements about technological artefacts refer to the fulfilment of technological function.

In chemistry just as in engineering, functional categories give rise to category-specified values. A chemist can talk about a "good oxidizer" or a "bad stabilizer", just as an engineer can talk about a "good car jack" or a "bad graphics card". Just as in engineering, the valuation criteria are defined by the function. If we know what an oxidizer is, then we know what a good oxidizer is – namely a substance that fulfils the function of an oxidizer efficiently. (In both cases, functions are usually considered contextually. A chemical that is a good oxidizer for one substance may not be so for another, less readily oxidized substance. Similarly, the tool that you call "a good hammer" when driving small tacks may not answer to that description when you drive 4 in. nails.)

However, in addition to these similarities there are also at least three important differences between the value discourses in the two disciplines.

First, in chemistry evaluation is more strictly limited to functionally defined entities than in technology. As was mentioned above, an engineer will have no difficulty in understanding phrases like "a good steel wire" or "an improved concrete pillar"; these phrases refer to the functions that steel wires and concrete pillars usually have. In contrast, chemists do not seem to use phrases such as "a good peroxide", not even in a context where it is obvious for instance that the substance in question will be used as an oxidizer. The chemist would then say: "This peroxide is a good oxidizer", not: "This is a good peroxide".

Secondly, whereas virtually all functional categories in technology seem to give rise to category-specified value statements, only some of the functional categories in chemistry give rise to such statements. In Sect. 13.4.1 we discussed two major classes of functional terms in chemistry: those denoting categories of substances and those denoting parts of molecules ("functional groups"). The latter do not seem to give rise to any category-specified value statements: There is no talk about good or bad alkyl groups.

Furthermore, the use of category-specified value statements is not universal even for the functional terms that denote categories of substances. Possibly the best counterexample is the notion of an acid. An acid is (according to the standard Brønsted-Lowry definition) a proton donor. This is clearly a functional definition. However, chemists do not refer to acids as "good" or "bad" but rather as "strong" or "weak". The strong acids, of course, are those that readily donate protons to other molecules.

There is an interesting difference here between the terminologies for transfers of electrons and protons. A molecule that efficiently increases the oxidation number of other molecules can be called a "good oxidizer", and one that efficiently reduces the oxidation number of other molecules a "good reductant", but as already mentioned we do not call a molecule that donates or receives protons efficiently a "good acid" respectively a "good base". The reason for this difference seems to be historical. The term "acid" goes back to antiquity and is thus much older than the current functional understanding of the concept of a proton donor (we can, however, say "a good proton donor"). The term "oxidizer" is much younger than "acid"; the oldest excerpt in the OED is from 1850 and refers to a functional property of the substance (namely the ability to convert another substance into an oxide; this account was replaced first by the general electron transfer account and later by the modern account in terms of oxidation numbers).

We can now summarize the first two differences between the uses of category-specified value statements in chemistry and in technology. Together they show that such statements have much wider application in technology than in chemistry. Category-specified value statements are made about virtually all classes of technological objects (including those that are not functionally defined), whereas they are only made about some of the functionally defined objects referred to in chemistry (namely most of the functionally defined classes of substances).

The third difference concerns the distinction between holistic and myopic evaluations of function. A functional evaluation is myopic if it only refers to the function *per se*; it is holistic if it also takes into account the surrounding circumstances that have an influence on whether the functionality in question can be made use of in practice. From a myopic point of view, a good oxidizer is a substance that readily increases the oxidation number of other molecules. A good oxidizer in this sense may nevertheless be useless in the laboratory or the chemical industry since it is too unstable or otherwise difficult to work with. From a holistic point of view, a good oxidizer is a substance that can be used efficiently in a laboratory or factory to oxidize other substances.

Chemists seem to vacillate between these two modes of speaking. As an example of this, a chemist may express the same state of affairs in either of these two ways:

> This is a good solvent for the substance, but we cannot use it since it is too explosive. (myopic value assignment)
>
> This is not a good solvent for the substance, since it is too explosive. (holistic value assignment)

Both modes of speaking seem natural; we can conclude that linguistic habits vacillate between the holistic and the myopic types of value assignment.

**Table 13.1** The use of instrumental value statements about artefacts in engineering and chemistry

| | Engineering | Chemistry |
|---|---|---|
| *Type of value statements* | Category-specified | Category-specified |
| *Valuation criteria* | Function-based | Function-based |
| *Classes of evaluated objects* | All functionally defined artefact classes. Some physically defined artefact classes | Many but not all functionally defined classes of chemical substances |
| *Holistic/ myopic* | Holistic | Holistic or myopic |

The same distinction is equally applicable to engineering. Consider an internal combustion engine that has a high energy conversion efficiency but tends to explode. This could be expressed in either of the following two ways:

> This is a good engine but we cannot use it since it tends to explode. (myopic value assignment)
> This is not a good engine since it tends to explode. (holistic value assignment)

However, the first of these expressions has a slightly absurd flavour. Most engineers would not express themselves in that way (or possibly do so tongue in cheek). Generally speaking, the holistic approach dominates in engineering, contrary to chemistry where neither of the two approaches seems to dominate over the other.

The similarities and differences between chemical and engineering practices in value assignments to artefacts are summarized in Table 13.1. How can we explain the differences? One reasonable hypothesis is that they depend on the relative importance of functional descriptions in the two disciplines. Chemistry and engineering both categorize their study objects in two ways, according to physical structure and according to function. But as already mentioned, functional descriptions have a more dominant role in engineering than in chemistry. There is an obvious historical reason for this, already indicated above in connection with the concept of an acid. Technology has always used functional descriptions. The functional descriptions in chemistry depend on modern understanding of molecular reactions, and are therefore of more recent origin.

As we have seen, the valuation of artefacts is closely bound to their functional characterizations. Therefore it is no surprise that technology, in which functional descriptions of artefacts have a greater role, also has a more extensive practice of assigning values to these artefacts. In this respect – and perhaps in others as well – chemistry seems to have an intermediate position between engineering and the other natural sciences.

## 13.5 Value-Assigning Artefacts

Finally, let us briefly consider the possibility of artefacts acquiring another role in relation to values: that of assigning values rather than having values assigned to them. Again, it is useful to distinguish between instrumental and moral values.

The use of technical artefacts to produce instrumental value assignments is commonplace. Machines are employed in many industries for quality control. Decisions to discard mass-produced objects due to manufacturing defects need not pass through a human head; it is often more efficient to design a machine to discard such objects automatically. This means that agency and decision-making have in a sense been transferred to a machine. In some cases the quality criteria programmed into machines may have a moral justification, for instance when potentially unsafe products are discarded. However, the machines operating according to these criteria would still not be called moral agents. A moral agent would presumably have to act upon its own moral reasons, not upon non-moral criteria that have been predetermined on the basis of moral reasoning by human beings.

It seems, therefore, as if ability to make one's own assignments of moral value is a necessary requirement for moral agency. We do not now know if technological artefacts can be designed that satisfy this requirement. But we do know that if they can, then they will be involved in the same type of complex value-based interrelations as humans are. The fact that this is at all a seriously considered possibility further confirms that technological artefacts have a greater potential than chemical artefacts for being associated in various ways with the realm of values.

## 13.6 Conclusion

Technology and chemistry both involve actions and activities that lead to the creation of new types of objects. In technology such activities are called "engineering design". In chemistry they are usually called "synthesis" (although the term "molecular design" is also popular). Largely because of the common design element there are important similarities between the uses of value statements in chemistry and technology. These similarities appear on two levels of our discourses on technology and chemistry.

First, value statements are made on what we may call the "macro level", referring to technological and chemical practices such as nuclear technology, biotechnology, or the use of pesticides or synthetic pharmaceuticals – or even to technology or chemistry in general. In public discussions on both technology and chemistry there is a noticeable abundance of negative appraisals on this level. There seems to be a close connection between such negative appraisals and the perceived novelty or unnaturalness of the products of technology and chemistry.

Secondly, value statements are also made on the "micro level" of specific technological artefacts or chemical molecules or parts of molecules. We distinguish between good and bad cars and also between good and bad oxidizers. As shown above, whereas the value statements on the macro level are associated with novelty, those on the micro level are strongly connected with functionality. When we describe a machine, a tool, or a catalyst as good or bad, then we refer to how well it fulfils the function assigned to it. Since functional talk has a more prominent role in technology than in chemistry, micro level value statements are also more common in technology.

The two types of value statements do not seem to be straightforwardly connected with each other, but they both need to be taken into account in studies of value assignments to humanly created objects. Further investigations of the similarities and differences between value assignments in technology and chemistry can contribute to our understanding of both disciplines.

**Acknowledgments** I would like to thank Margareta Blomberg, Peter Kroes, Joachim Schummer, and Peter-Paul Verbeek for valuable comments on an earlier version of this paper.

## References

Anderson-Rowland, M. R. (1996, November). A first year engineering student survey to assist recruitment and retention. In *Proceedings, FIE conference*, Salt Lake City, Utah (pp. 372–376).

Becker, F. S. (2009, July 1–4). Why not opt for a career in science and technology? An analysis of potentially valid reasons. In M. v. d. Bogaard, E. d. Graf, & G. Saunders-Smits (Eds.), *Proceedings of 37th annual conference of SEFI. Attracting young people to engineering. Engineering is fun!*. Rotterdam. http://www.sefi.be/wp-content/abstracts2009/Becker.pdf

Berdonosov, S. S., Kuzmenko, N. E., & Kharisov, B. I. (1999). Experience in chemical education in Russia: How to attract the young generation to chemistry under conditions of 'Chemophobia'. *Journal of Chemical Education, 76*, 1086–1088.

Bunge, M. (1966). Technology as applied science. *Technology and Culture, 7*, 329–347.

Bunge, M. (1988). The nature of applied science and technology. In V. Cauchy (Ed.), *Philosophy and culture, proceedings of the XVIIth congress of philosophy* (Vol. II, pp. 599–604). Montréal: Éd. Montmorency.

Hansson, S. O. (2003). Are natural risks less dangerous than technological risks? *Philosophia Naturalis, 40*, 43–54.

Hansson, S. O. (2006a). Category-specified value statements. *Synthese, 148*, 425–432.

Hansson, S. O. (2006b). Defining technical function. *Studies in History and Philosophy of Science, 37*, 19–22.

Hansson, S. O. (2006c). A note on social engineering and the public perception of technology. *Technology in Society, 28*, 389–392.

Hansson, S. O. (2007). What is technological science? *Studies in History and Philosophy of Science, 38*, 523–527.

Heilbronner, E., & Wyss, E. (1983). Bild einer Wissenschaft: Chemie. *Chemie in unserer Zeit, 17*, 69–76.

Kroes, P. (2006). Coherence of structural and functional descriptions of technical artefacts. *Studies in History and Philosophy of Science, 37*, 137–151.

Kroes, P., & Meijers, A. (Eds.). (2006), The dual nature of technical artefacts (special issue). *Studies in History and Philosophy of Science,* 37, 1–158.

Mitcham, C., & Schatzberg, E. (2009). Defining technology and the engineering sciences. In A. Meijers (Ed.), *Handbook of the philosophy of science: Vol. 9. Philosophy of technology and engineering sciences* (pp. 27–63). Amsterdam: Elsevier.

Rawling, P. (1990). The ranking of preference. *Philosophical Quarterly, 40*, 495–501.

Read, D. (2010). A U.K. approach to counter declining enrollment in chemistry and related disciplines at the university level. *Journal of Chemical Education, 87*, 898–900.

Rescher, N. (1968). *Topics in philosophical logic*. Dordrecht: Reidel.

Schummer, J. (1997). Challenging standard distinctions between science and technology: The case of preparative chemistry. *Hyle, 3*, 81–94.

Schummer, J. (2001). Ethics of chemical synthesis. *Hyle, 7*, 103–124.

Schummer, J. (2003). The notion of nature in chemistry. *Studies in History and Philosophy of Science, 34*, 705–736.

Sjöberg, L. (2002). Attitudes toward technology and risk: Going beyond what is immediately given. *Policy Sciences, 35*, 379–400.

Tertullianus, Q. S. F. (ca. 200). *De cultu Feminarum*. Available on http://www.intratext.com/IXT/LAT0750

Vermaas, P. E., & Houkes, W. (2006). Technical functions: A drawbridge between the intentional and structural natures of technical artefacts. *Studies in History and Philosophy of Science Part A, 37*, 5–18.

von Wright, G. H. (1963). *Varieties of goodness*. London: Routledge & Kegan Paul.

Yurtseven, H. O. (2002). How does the image of engineering affect student recruitment and retention? A perspective from the USA. *Global Journal of Engineering Education, 6*(1), 17–23.