Computational Social Sciences

Maria Xenitidou
Bruce Edmonds
Editors

The Complexity of Social Norms

Springer

Computational Social Sciences

A series of authored and edited monographs that utilize quantitative and computational methods to model, analyze, and interpret large-scale social phenomena. Titles within the series contain methods and practices that test and develop theories of complex social processes through bottom-up modeling of social interactions. Of particular interest is the study of the co-evolution of modern communication technology and social behavior and norms, in connection with emerging issues such as trust, risk, security, and privacy in novel socio-technical environments.

Computational Social Sciences is explicitly transdisciplinary: quantitative methods from fields such as dynamical systems, artificial intelligence, network theory, agent-based modeling, and statistical mechanics are invoked and combined with state-of-the-art mining and analysis of large data sets to help us understand social agents, their interactions on and offline, and the effect of these interactions at the macro level. Topics include, but are not limited to social networks and media, dynamics of opinions, cultures and conflicts, socio-technical co-evolution, and social psychology. Computational Social Sciences will also publish monographs and selected edited contributions from specialized conferences and workshops specifically aimed at communicating new findings to a large transdisciplinary audience. A fundamental goal of the series is to provide a single forum within which commonalities and differences in the workings of this field may be discerned, hence leading to deeper insight and understanding.

Maria Xenitidou • Bruce Edmonds
Editors

The Complexity of Social Norms

Springer

Editors
Maria Xenitidou
Department of Sociology
University of Surrey
Guildford, Surrey, UK

Bruce Edmonds
Centre for Policy Modelling
Manchester Metropolitan University
Manchester, UK

ISBN 978-3-319-05307-3 ISBN 978-3-319-05308-0 (eBook)
DOI 10.1007/978-3-319-05308-0
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014939118

Printed on acid-free paper

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

Contents

Chapter 1
The Conundrum of Social Norms

Maria Xenitidou and Bruce Edmonds

Motivation

Consciously or otherwise people decide what to do bearing in mind what they think is acceptable/unacceptable to others around them. These standards of acceptability can be called social norms. Thus, the idea of social norms lies at the heart of sociology—how individual behaviour is constrained by (the individual's view of) the expectations of others. There is often considerable agreement between participants as to when a social norm is violated, and people report that what they perceive as social norms impact upon them both in thought and action. Some are bold enough to call social norms "the grammar of society" (Bicchieri, 2005).

However, simplistic conceptions of social norms are plagued with difficulties. Their independent existence as reified entities to be labelled and tracked is problematic. What seems obvious to all about what a social norm is tends to dissolve upon closer examination. What is acceptable or not seems very changeable according to the time, place and social context of any action. They critically rely on the perceptions of individuals, and yet accounts of norms as only conventions are insufficient to explain their persistence. For all these reasons (and more) norms have become problematic to study and so, in the last 20 years, have been relatively neglected.

Despite all these difficulties, however, what we call social norms clearly have both social efficacy and a high level of inter-subjective reality. In Chap. 4, Chris Goldspink gives an example of a person arriving in the UK from New Zealand and trying to start a conversation with strangers waiting at a bus stop. Disapproval of this innocent action

M. Xenitidou (✉)
Department of Sociology, University of Surrey, Guildford, Surrey, UK
e-mail: M.Xenitidou@surrey.ac.uk

B. Edmonds
Centre for Policy Modelling, Manchester Metropolitan University, Manchester, UK

M. Xenitidou and B. Edmonds (eds.), *The Complexity of Social Norms*, Computational Social Sciences, DOI 10.1007/978-3-319-05308-0_1,

was immediately apparent—whatever a social norm *is* it has force, in this case stopping a New Zealander from trying to start casual conversations with people he does not know. Second, when that person talked to others about his experience they all reported that what happened was normal for the UK, indeed expected—whatever a norm *is*, it derives from a near unanimous recognition across a whole group of people (even those who don't hold with the norm recognise its existence). This combination of social efficacy with widespread inter-subjective recognition gives social norms a *reality* that is in stake contrast to the difficulties in identifying and studying them. Ultimately, one cannot pass judgement upon the meaningfulness of a phenomena's practical existence on the basis of the difficulty of its identification—just because disease is spread in complex ways by complex organisms that are difficult to detect does not mean that "disease" is not a meaningful idea[1] or that disease is not real.

This book does not flinch from the complexity of the phenomena it is interested in. It brings together a disparate set of authors, each of whom accepts the *reality* of social norms in different ways but who also seek to explore their complexity. Part of this complexity lies in the way that what is recognised and identified as social norms is an abstraction of a complex and dynamic interaction of many aspects and levels. Social norms emerge and dissolve over time and within different groups of people both cognitively and socially, so it is these complexities that this book investigates. In this way this book aims to play a part in revitalising interest in social norms by taking a complex and dynamic perspective, replacing a static picture of norms as social *objects* with norms associated with a socially recognisable complex—an intertwined set of cognitive and social processes partially locked in by emergent and immergent forces.

This viewpoint reflects some of the concerns of "complexity science". Usually, "complexity science" has emphasised a purely bottom-up approach, whereby complex phenomena might result from the interaction of simpler parts—in other words, they *emerge*. However, for social scientists this is only half of the picture, with the other half being how society constrains the actions of individuals—what has been called immergence or "downward causation" (Campbell, 1974). Although many ideas from complexity science have been applied to social phenomena in an over-simple and reductionist manner, recent developments have resulted in the beginning of a synthesis with other social science approaches. So that richer and more descriptive approaches to social phenomena are used along with dynamic approaches. This goes to the heart of social science because it allows explicit explorations of how interaction at a micro-level leads to the emergence of macro-phenomena and how the macro-level, societal trends and institutions can act back upon the micro-level interactions.

Thus, this book does not restrict itself to the views that derive from "complexity science". Rather, it seeks a new alignment, where many processes and mechanism are reconsidered, under the umbrella label of "social norms" inspired and informed by many developments including "complexity science". In this way this important set of phenomena can, once again, play a central part in the understanding of society,

[1] Of course, in an age before appropriate tools to enable effective study of such phenomena it might mean that one decides that it is not feasible to attempt a study of it. Indeed, it might be that there are *many* cases where the identification of social norms is problematic, but these do not make them unreal.

albeit in a complex, dynamic and context-dependent manner. This volume collects together a variety of different approaches to norms, all of which go beyond simplistic or static pictures of social norms, but rather as: constantly changing, shifting over time and socio-cultural contexts, both appearing and being passed over. The volume aims to re-invigorate the study of norms and normative behaviour by allowing these complexities back into the picture.

The Issues

This book starts from the assumption that normative behaviour—behaviour that is characterised by its adherence to established standards of correctness and propriety—is integral to all "cultures" or "folk ways". Normative behaviour is revealed in everyday discourse, for example in the negotiation of antithetical norms (e.g. providing for one's family by stealing versus committing a crime), the bewailing of the breakdown of social norms (e.g. when someone jumps a queue at a bus stop) or in seeking to establish new norms (e.g. drinking and driving, or speeding).

This book explores the view that normative behaviour is a part of a complex of social mechanisms, processes and narratives that are constantly shifting. From this perspective, norms are not a kind of self-contained social object or fact, but rather an interplay of many things that we label as norms when we "take a snapshot" of them at a particular instant. Further, this book pursues the hypothesis that considering the emergent and dynamic aspects of these phenomena sheds new light on them.

The sort of issues that this perspective opens to exploration include:

- Under what circumstances, what combination of processes and factors will result in something we call a social norm?
- How do new social norms emerge and what kind of circumstances might facilitate such an appearance?
- When do existing social norms lose their power, becoming formalised, empty or simply ignored?
- To what extent are social norms linked to particular groups or societies?
- How context-specific are the norms and patterns of normative behaviour that arise?
- How does the cognitive and the social aspects of norms interact over time?

How Have These Questions Been Approached by Different Disciplines

Social norms have primarily pre-occupied sociology, psychology, economics, politics, international relations law and—to a lesser extent in recent times—philosophy. In sociology, the decline of the influence of functionalism saw a parallel decline in discussions and work on social norms. Michael Hechter and Karl-Dieter Opp (2001) who made this observation, focused on norm emergence subscribing to the

instrumental theory of the emergence of norms. Their book includes reviews of existing theory and research on social norms in sociology, law, economics and game theory and focuses on the emergence of norms from the perspectives of: institutionalism and individualism, social networks, evolutionary psychology and behaviour-based and externality-based explanations. However, an evaluation of the developments in the study of norms is missing as no associations or classifications based on underlying criteria are made other than the topic of *norm emergence* itself and a general emphasis on an *instrumental* view of norms. Christina Bicchieri (2005) has focused on norms as a system of rules which are not written but which are implicit in the operations of society and define society and the way in which human groups live. The emergence of, adherence to and demise of social norms are seen from a *socio-cognitive* perspective, and a *game theoretic* approach is employed to capture the dynamics of these processes. Bicchieri (2005) places crucial emphasis on the definition and classification of norms; informal norms are classified into: social norms, conventions and descriptive norms. Norms appear to be treated as self-fulfilling expectations visualised in coordination games taking into account how situations are categorised and which scripts are subsequently activated. Therefore, context and situated meanings are taken into account and explored experimentally using the Ultimatum, Trust, Dictator and Social Dilemma games. Thus, although the Bicchieri (2005) examines the dynamics of the emergence of, adherence to and demise of social norms, her work has focused on a single perspective and a single approach—social cognition and game theory, respectively (cf. Chap. 3).

In psychology, the study of social norms has remained within a cognitive perspective, which has hindered broader attempts to conceptualise social norms (see, for example Raz, 1999; Terry & Hogg 2000; Dubois 2002 cf.; Howitt et al., 1989). Raz (1999) analyses the role of reason and exclusionary rules, "paving the way to a unified normative account". *Games* are used as to exemplify normative systems, and the analysis extends to some aspects of normative discourse. Thus, this looks at the roots of normative reasoning mostly from the *individual* point of view. Although the book touches on dynamic aspects, it is primarily *structural* in its approach. Terry and Hogg (2000) bring together attitude researches on how the social context in the form of social norms and group membership may influence attitudes. The book emphasises a *socio-cognitive* perspective and includes research in developmental psychology, self-identity perspectives, social identity theory and self-categorisation theory, cognitive dissonance theory and a connectionist approach to cognitive modelling. While there is an emphasis on context (see also Bicchieri, 2005), it is restricted to the socio-cognitive perspective. Dubois' collection (2002) focuses on how behaviours are socially regulated, starting from the premise that norms not only affect what we do but also how we think and the judgments we make. The collection seeks to establish that the *social judgment norm construct* and the *socio-cognitive* approach in which it is embedded explains social thinking in diverse contexts. The current volume is different from this in its emphasis on the emergence of normative patterns within a fundamentally dynamic approach.

Whilst philosophy seems to have abandoned discussions and work on norms (cf. Critto (1999) who discusses the scales that societies use to effect social change,

focusing on Argentinean society), economics, politics, international development and law have engaged in discussions and scholarly activity on norms, but from within their own subject area. For example, Posner (2002) looks at the relationship of law to social mechanisms such as norms, asking what the role of law in a society in "which order is maintained mostly through social norms, trust and non-legal sanctions" might be. Thus, it looks at how the law might support or undermine social norms. This work considers many aspects of life from the perspective of the impact of laws, including game-theoretic approaches, but does not take a fundamentally dynamic view (Posner 2007 covers the same area between *law and norms*). Hetcher (2002) looks at the role of laws across the Internet, particularly those to do with privacy and tort law. It again uses a *game-theoretic* framework. Perkins (2003) has focused on interventions using social norms to try and reduce substance abuse by young people. Juètting et al. (2007) consider how social norms might hinder or help the development of countries with weak institutional structures. Platteau's earlier book (2000) on the same subject draws on his fieldwork, arguing that *norms and institutions* are shaped by a complex of physical and social conditions. Finally, the study of social norms has attracted complexity and computer scientists. For example, Christina Bicchieri, Richard Jeffrey and Brian Skyrms (1997) have taken a dynamic approach from a largely *game-theoretic* perspective, looking at some iterated prisoner's dilemma games and similar simulations and analyses.

Overall, the account above indicates a strong concentration of studies on norms from instrumentalist, socio-cognitive and game-theoretic perspectives. It is for this reason that we feel re-starting the discussion on social norms from the dynamic, complexity viewpoint is needed, bringing different methodological and theoretical perspectives together as well as theoretical and methodological discussions on norms from a variety of substantive areas. These include: philosophy and sociology, especially new epistemologies and methods emphasising a processual view of social phenomena; social psychology, especially the study of social and group influence processes; computational and institutional economics, especially focusing on the processes of self-organisation; politics and international relations, especially on the processes of social order and control; criminology and law; computer—including artificial intelligence (AI)—and complexity sciences and new epistemologies and methods of studying norms; simulation approaches and/or the combination of simulation methods with other methods and, overall, the social effects of social norms, why they might appear or disappear.

The Structure of the Book

The key idea of this book is to show how a dynamic and complex approach to social norms is inherent in a number of different developing approaches. Thus, the core of the book is a collection of chapters describing these approaches allowing commonality between these approaches to be clear. This core is framed by an introduction and some synthetic critical pieces reviewing these, drawing out the synergies, compatibilities and differences.

The current, first chapter is an introduction to the book. Chapter 1 sets the scene by reviewing the history of thought about norms in the social sciences, arguing for the centrality of norms, but as an emergent phenomena resulting from underlying dynamic and complex phenomena.

Parts I & II constitute the core of the book. These consist of a series of approaches to the study and understanding of norms each coming from a different direction and tradition, but all taking a new view of norms as an umbrella terms for a set of complex social and individual phenomena.

Part I: *The Complex Roots of Social Norms* includes five chapters that emphasise different perspectives that unearth some of the sources and reasons for the nature and complexity of social norms. These viewpoints into the complexity of norms are far from contradictory, but exactly how all these pieces fit together is not entirely clear, leaving some room for subtle tensions between these contributions.

First, Wesley Perkins (Chap. 2) registers a case for a dynamic view to norms by discussing the extent to which group norms are misperceived by group members and the implications of this perceptual error—"reign of error" as he calls it—for personal actions that are presumed to be influenced by norms. The theoretical case draws on extensive empirical research on peer risk behaviour norms among youth and young adults. The chapter aims to establish a link between perceptions, attitudes and behaviour positing the "problem" of misperceptions as one of the reasons why norms are dynamic—the gap between perception of self and others in terms of norms of behaviour as the drivers behind norm lock-in, change and intervention.

In Chap. 3, Cristina Bicchieri and Hugo Mercier argue for a relationship between norms and beliefs and introduce the notions of discussion and deliberation as the means to achieve change through arguments. They start from the premise that social norms—behavioral rules supported by a combination of empirical and normative expectations—play an important role in both explaining and changing negative practices. Norm change or the creation of new norms can be effected by acting upon empirical expectations—our belief(s) of what should be done in a given situation—and normative expectations—our belief(s) of others' beliefs of what should be done in a given situation and, then, by introducing mechanisms—discussion and deliberation—that will bridge expectations and behaviour.

In Chap. 4, we move on from interventionist accounts, but keeping in line with communication, Chris Goldspink presents an emergentist viewpoint norms from the perspective of an enactive approach to cognition. The chapter reviews extensive literature making some fundamental criticisms to accounts that only address the micro- or macro-level of explanation. Instead, it develops a level-based account of emergence which considers the defining features of human social agents –"critical cognitive capabilities" such as: affect and emotion, agency, consciousness, self-awareness, identity, cultural tools and language–significant to normative behaviour. Finally, the enactive approach to cognition–enactment of structural coupling among unities which are self-aware and linguistically capable in the environment they enter–lays out the role of these human cognitive capabilities and the ways in which they interact.

Corinna Elsenbroich (Chap. 5) introduces the notion of "we-intentionality" or "shared intentionality" to the study of normative behaviour. In particular, the

chapter criticises sociology for adapting the prism of individualism and argues that the unique feature of humans which brings about this unique social world is "we-intentionality"–that human beings do not only behave following their own intentions but are unique in joining intentions with others. Thus, assuming the ability to share intentions enables modelling complex forms of normative behaviour, such as institutions and culture as well as the dynamics of normative systems. The author discusses ways in which we-intentionality might be operationalised in agent-based models of normative behaviour.

In Chap. 6 Christine Horne bases the relational foundation of norms on dependence. The chapter focuses on social relationships as the key factor in explaining norm enforcement and, in particular, on (inter)dependence amongst group members—the extent to which one values their relationship with others as well as the goods that he or she can get from that relationship. Thus, following norms and sanctioning non-followers ("deviants") depends on whether group members are dependent on one another and value their relationship(s) and on the extent of this dependence. The author offers evidence from a series of laboratory experiments which support these theoretical claims.

Part II: *Methods and Epistemological Implications of Social Norm Complexity* includes three chapters which are centred more around methodological considerations (e.g. Agent-based Modelling).

Here, Brigitte Burgemeestre, Joris Hulstijn and Yao-Hua Tan (Chap. 7) make a case for norm change and emergence focusing on the concept of "open norms" used in regulatory compliance and exploring it through a specific case study—that of the regulations concerning kilometre registration for lease car drivers in the Netherlands. Open norms refer to norms which leave room for contextual interpretation about how they should be implemented, thus leading to (some kind of) norm emergence. The authors compare their findings to relevant literature from the multi-agent systems (MAS) field and suggest ways to extend MAS research on norm emergence.

In Chap. 8, Giulia Andrighetto, Daniel Villatoro and Rosaria Conte focus on norm dynamics and cognition viewing social norms as guides of conduct transmitted from one agent to another through normative requests or evaluations. The authors present a multilevel model to show the mental path followed by a norm in regulating human behaviour and to specify the cognitive "ingredients" and processes necessary for a normative request to be complied with.

Finally, Marco A. Janssen and Elinor Ostrom (Chap. 9) explore the consequences of visibility on behaviour—in other words what happens to norm following in the situation of having incomplete information about the collective action in which these norms make sense. The authors explore this by developing an agent-based model that describes a population of agents who share a common-pool resource, have a norm regarding when to harvest from the resource and varying levels of visibility of others' actions. Their results suggest that transparency and complete information are necessary in order to maintain norms that enhance sustainable use of commons.

The book ends with *Part III*: *Evaluating Complex Approaches to Norms*, which consists of two chapters reviewing and reflecting up the approaches in Parts I & II, commenting upon them and providing a synthetic critique.

In the first chapter, Flaminio Squazzoni (Chap. 10) focuses on two main aspects in the contributions of the volume: social context and cognition. He emphasises the need to operationalise social context in specific terms, by considering, in particular, the ways in which social structure influences behaviour. Second, he discusses how both purely cognitive models per se and the use of experiments and simulation only are insufficient in understanding the social norms puzzle.

In the second, Bruce Edmonds (Chap. 11) identifies three difficulties of understanding social norms due to their nature. These are that: norms simultaneously involve many levels (e.g. cognitive and social); are dynamic, continuously emerging, changing and falling into disuse and are highly context-dependent with different norms pertaining to different situations, identities and social groupings. The consequences of these three difficulties are discussed in turn, drawing out how the different chapters in this volume recognise and deal with them. Some tentative conclusions as to some ways forward for the study of norm-constrained behaviour are suggested.

Acknowledgments The editors acknowledge support from the EU 6th framework project, EMergence In the Loop: simulating the two way dynamics of norm innovation (EMIL), contract number 033841, from which this book emerged.

References

Bicchieri, C. (2005). *The grammar of society*. Cambridge: Cambridge University Press.
Bicchieri, B., Jeffrey, R., & Skyrms, B. (1997). *The dynamics of norms*. Cambridge: Cambridge University Press.
Campbell, D. T. (1974). 'Downward causation' in hierarchically organized biological systems. In F. J. Ayala & T. Dobzhansky (Eds.), *Studies in the philosophy of biology* (pp. 179–186). London: Macmillan Press.
Critto, A. (1999). *Choosing models of society and social norms: improving choices and quality of life*. Lanham, MD: University Press of America.
Dubois, N. (2002). *Sociocognitive approach to social norms*. London: Routledge.
Hechter, M., & Opp, K.-D. (2001). *Social norms*. New York: Russell Sage.
Hetcher, M. (2002). *Norms in a wired world*. Cambridge: Cambridge University Press.
Howitt, D., Billig, M., Cramer, D., Edwards, D., Kniveton, B., Potter, J., et al. (1989). *Social psychology: conflicts and continuities*. Milton Keynes/Philadelphia, PA: Open University Press.
Juètting et al. (2007). Informal institutions: How social norms help or hinder development (OECD; Development Centre studies).
Perkins, W. (2003). *The social norms approach to preventing school and college age substance abuse: a handbook for educators, counselors and clinicians*. New York: John Wiley.
Platteau, J.-P. (2000). *Institutions, social norms and economic development*. London: Routledge.
Posner, E. (2002). *Law and social norms*. Cambridge, MA: Harvard University Press.
Raz, J. (1999). *Practical reason and norms*. Oxford: OUP.
Terry, D. J., & Hogg, M. A. (2000). *Attitudes, behaviour and social context*. Hillsdale, NJ: Lawrence Erlbaum.

Part I
The Complex Roots of Social Norms

Chapter 2
Misperception Is Reality: The "Reign of Error" About Peer Risk Behaviour Norms Among Youth and Young Adults

H. Wesley Perkins

Introduction

Social norms were viewed as the cultural and structural underpinnings of human behaviour and organization and were a key focus in the founding of the discipline of sociology as exemplified in the classic theory and research of Emile Durkheim. In addition to the study of how widely held beliefs and widely practised behaviours ground individual actions and provide people with a sense of meaning and purpose, over half a century of voluminous empirical studies in social psychology point to the power of group norms in influencing individual action. These experiments date all the way back to the classic experiments of Solomon Asch (1951, 1952, 1956) and Musafer Sherif (1936, 1972). Numerous topics remain for contemporary study, however, regarding the complexity of how social norms are constructed (or emerge and evolve) and how they exert control over individuals' behaviour.

In this chapter I focus on a particular theoretical and empirical issue that has emerged in recent decades, that being the extent to which group norms might be misperceived by group members and the implications of this perceptual "error" for personal actions that are presumed to be influenced by norms. On the one hand, actual group standards may exist that control or influence individual behaviour as a contextual effect, regardless of one's consciousness of a particular norm. On the other hand, people may behave in accordance with what they perceive to be peer group standards and also attempt to influence the behaviour of others to act in line with their normative perceptions, irrespective of the accuracy of these perceptions.

H.W. Perkins, Ph.D. (✉)
Department of Anthropology and Sociology, Hobart
and William Smith Colleges, Geneva, NY, USA
e-mail: perkins@hws.edu

M. Xenitidou and B. Edmonds (eds.), *The Complexity of Social Norms*,
Computational Social Sciences, DOI 10.1007/978-3-319-05308-0_2,

Furthermore, I specifically focus this theoretical discussion and literature review of misperceived norms on one broad topic area of applied research, that being norms regarding risk behaviours among youth and young adults. The rationale for concentrating on this area of research in my examination is straightforward. Although a few studies regarding other topics have appeared on occasion examining misperceived norms, one of the earliest empirical investigations was focused on youth risk behaviour (student alcohol abuse) and it simultaneously suggested an approach for applying the model to address this widely acknowledged social problem (Perkins & Berkowitz, 1986). From that initial study to the present, by far the largest body of empirical studies on misperceived norms has been devoted to research on youth and young adult risk behaviours. This area of research now provides enough collective studies to be able to generalize about misperceived norms in this area and the conclusions drawn have direct implications for promoting health and well-being.

I initially review the social science research empirically demonstrating substantial discrepancies in actual and perceived norms concerning risk behaviour. I then consider research on the empirical correlation of perceived norms with personal behaviour as well as research on that association independent of and in comparison to the association between actual norms and personal behaviour across populations. Finally, I review theory and research literature examining what produces these misperceptions, whether misperceptions can be altered or corrected by revealing accurate peer norms within the social group, and whether any change achieved in perceived norms produces subsequent change in individual behaviour.

This chapter focuses on this set of questions as one way in which norms may be "dynamic." That is, actual youth and young adult norms regarding healthy and risky behaviours may be more or less influential upon individuals depending on how these norms are filtered through the individuals' perceptual assessments and interpretations of peer norms. If perceived norms are a salient aspect of normative influence, to the extent that perceptions of norms can be changed, the outcome of such change in perceptions may be a concomitant shift in personal attitudes and behaviours.

At the outset of any discussion on social norms one must acknowledge that the search for a specific definition of social norms has not produced consensus (Horne, 2001). Various definitions concentrate on sanctions, values ("oughtness"), or behavioural regularities (Hechter & Opp, 2001). Some social scientists restrict the definition to social expectations that are clearly backed by rewards and consequences to assure widespread compliance while others focus on particular attitudes or beliefs that implicitly, if not explicitly, convey beliefs about morally acceptable behaviour. Other theorists and researchers focus on the instrumentality of social norms and point to shared practices and beliefs that function to bind people together in solidarity and provide a unified identity for the group. Still others adopt a broad empirical approach by examining the most common or majority attitudes in a group (injunctive norms) and the most common or majority behaviours in a group (descriptive norms) (Cialdini, Reno, & Kallgren, 1990) and how they impact individual attitudes and behaviours as well as group functioning. Recognizing that definitional matters can be important but also that resolution of the differences in definition is not likely

or essential for the discussion that follows, the latter broad definitional approach—simply identifying norms as the dominant attitudes (injunctive norms) and practices (descriptive norms) of a group—is adopted here.

Actual Norms and Perceived Norms

Few social scientists would disagree with the claim that conformity to peer group norms is a widespread phenomenon and that peer influence, in addition to personal attitudes, is a powerful determinant of personal actions in many group contexts as individuals look to others in their midst to help define the situation and give guidance on expected behaviours. Indeed, although many people frequently think of themselves as individuals in their actions, a considerable degree of peer influence is consistently documented in laboratory experiments, social surveys, and observations of crowd behaviour. In studies on antecedents of personal health-related behaviours, for example, extensive evidence has supported the theory of reasoned action (Ajzen & Fishbein, 1980) and its extension, the theory of planned behaviour, which posits norms as a determinant of personal behaviour along with personal attitudes and perceived behavioural control (Ajzen, 2001, 2002; Ajzen & Madden, 1986).

Most research exploring the potential influence of social norms on personal behaviour has failed to distinguish, however, between the potential influence of actual group norms and the perception of norms. The research literature on normative influence prior to the mid-1980s provides many studies that (1) examine the effects of variation in aggregate group characteristics on individual attitudes and behaviours but do not consider perceived norms, or (2) use subjective assessments of peer norms as a proxy for actual norms when predicting the effect of norms on personal behaviour without directly considering the accuracy of these subjective reports of peer norms. Systematic examination about the question of accuracy of perceived peer norms and the subsequent empirical question about the simultaneous relative influence of both actual and perceived norms has emerged only in the last few decades (Perkins, 2003a). Here, one finds the most detailed theoretical explications and reviews of the most extensive empirical research (Berkowitz, 2005; Borsari & Carey, 2001; Carey, Borsari, Carey, & Maisto, 2006; Perkins, 1997, 2002, 2003b) concentrating primarily on alcohol and substance abuse among adolescents and young adults.

The Pervasiveness of Misperceived Peer Norms

The first study to bring concentrated attention to misperceived norms by examining the possible systematic discrepancy between actual peer norms (as reflected in the aggregate of reported personal attitudes and behaviours) and perceived norms was

focused on high-risk drinking among university students at one small institution of higher education in the USA (Perkins & Berkowitz, 1986). Large discrepancies were uncovered in that study between what was most typical of students' attitudes and behaviours and what was perceived to be most typical. Most students misperceived the norm by substantially overestimating the permissiveness of peer drinking attitudes and the extent of alcohol consumption. Students did so even though actual drinking norms were relatively heavier than what is found in many collegiate settings, due to the school's socio-demographic characteristics and regional setting. As part of the survey, students were given a range of five possible responses to indicate their attitudes toward alcohol use from the most conservative (drinking is never good) to the most permissive (frequent intoxication is acceptable and even if it interferes with other responsibilities). About 14 % held a relatively conservative personal attitude, about 66 % took a moderate position, and about 19 % were relatively permissive believing that frequent intoxication or intoxication that occasionally interfered with academics and other responsibilities was acceptable (only 1 % did not respond to the question). Thus, the vast majority of responses—and hence the norm for personal attitudes—was shown to be moderate. But when asked to give their impression of the general campus norm in the same survey, students painted a very different picture. Using identical response categories, virtually no one perceived the general norm to be conservative, only about one-third perceived it as moderate (the actual norm), and almost two thirds (63 %) saw their peers on campus as having a very permissive attitude toward drinking. Thus, while four-fifths of students believed that one should never drink to intoxication or that intoxication was acceptable only in limited circumstances, almost two-thirds thought their peers most typically believed frequent intoxication or intoxication that did interfere with academics and other responsibilities was acceptable.

This gross misperception of drinking norms was not simply the result of a particular historical situation momentarily distorting students' perceptions. Research conducted at multiple time points several years later at the same institution demonstrated the same pattern of drinking norm misperceptions (Perkins, 1994). Moreover, following the initial study, a similar pattern of dramatic misperceptions about peer drinking norms was subsequently found to exist in studies of a variety of other individual colleges and universities in the USA. For example, students at a New England state university (Burrell, 1990) perceived their friends as heavier drinkers than themselves, and among students attending a large university in the Northwest (Baer & Carney, 1993; Baer, Stacy, & Larimer, 1991), misperceptions of peer drinking norms were found to persist across gender and housing types. Page, Scanlan, and Gilbert (1999) also found that both males and females overestimated the extent of heavy drinking among peers of the same and opposite gender at a school in the Northwest. In survey investigations using multiple strategies, Prentice and Miller (1993) found misperceptions of peers' attitudinal norms about drinking among students at a prestigious east coast private university. Misperceptions of frequent or heavy episodic drinking were uncovered in a midsized Midwestern state university (Haines & Spear, 1996), a large state university in the Southwestern USA (Johannessen & Glider, 2003) and a midsized public university in the Mid-Atlantic

East coast region (Jeffrey, Negro, Miller, & Frisone, 2003). Research on specific behaviours such as preparty drinking and drinking game participation has also revealed substantial overestimates of the peer norm (Pedersen & LaBrie, 2008).

Although most research on misperceived norms has focused on student drinking, the phenomenon is not uniquely characteristic to the consumption of alcohol, but extends to other risk behaviours. For example, Hancock and Henry (2003) found that while the past month prevalence of smoking tobacco was between 30 and 40 % for two large public universities in the southeastern USA, students on average estimated the prevalence among peers to be 54 and 57 % at these schools. Although abstinence from marijuana use was the norm for three northwestern colleges, Kilmer et al. (2006) found that students grossly misperceived the norm with 98 % believing that the students in general used marijuana at least once per year if not more frequently. LaBrie, Hummer, Lac, and Lee (2010) have similarly reported that students misperceive injunctive (attitudinal) peer norms about marijuana. Another study conducted at one large university found 70 % of students overestimating peer use of non-medical prescription stimulants and prescription opioids (McCabe, 2008).

In a nationwide study of over 45,000 students attending 100 colleges and universities in the USA, Perkins, Meilman, Leichliter, Cashin, and Presley (1999) found a consistent difference between the self-reported frequency of drinking and students' perceptions of the frequency of peer alcohol consumption in campus contexts where abstinence or infrequent use were the median of self-reports and also where the median of self-reports revealed more frequent actual use. Furthermore, students in this study substantially overestimated the frequency of peer use of tobacco, marijuana, cocaine, amphetamines, sedatives, hallucinogens, opiates, inhalants, designer drugs, and steroids. A subsequent nationwide study of over 72,000 students attending 130 schools across the USA (Perkins, Haines, & Rice, 2005), likewise, found a consistent pattern of misperceptions among students across all types of institutions when examining the quantity of alcohol consumed, regardless of variation in the actual norm across schools. Although actual norms for the number of alcoholic drinks consumed at parties and social occasions ranged from abstinence for a few schools to a high of seven drinks in one institutional setting (with norms ranging from two to five drinks in most school settings), the majority of students attending schools with each level of actual consumption substantially overestimated the consumption of local peers.

When this consistent evidence of dramatic misperception is presented, a question often arises concerning the possibility that individuals may be simply underreporting their own behaviour rather than misperceiving the norms of peers. Several arguments counter this possibility, however. First, the survey evidence reported here is almost all gathered in anonymous surveys, thus reducing presumed pressure to hide personal behaviour. Second, large gaps between actual norms based on self-report and perceived norms are found in circumstances where the behaviour is legal (e.g. tobacco use and alcohol use in young adult populations) in addition to research on illegal behaviour. Third, these large misperception gaps with actual norms are also found based on questions about personal attitudes and perceived attitudes of others which dismisses the notion that the gap could simply result from a bias in

recall error in self-reported behaviour. Fourth, theoretical logic and research about normative influence would suggest that any bias in self report would operate in the direction of minimizing the gap between self-reported attitudes/behaviours and perceptions of the norm. Fifth, research based on breath analyzer studies to determine actual drinking norms rather than relying solely on aggregated self-reports (e.g. Foss, Marchetti, & Holladay, 2001; Thombs, Olds, & Snyder, 2003) also supports the finding that students typically perceive the norms for the amount of drinking among peers to be substantially greater than is actually the case, and that they do not, on average, under report their own consumption.

In recent years findings of pervasive misperceptions of alcohol and drug use norms among university students have also been documented in several studies outside the USA (McAlaney, Bewick, & Hughes, 2010). For example, in a study of students attending a large university in New Zealand, Kypri and Langley (2003) found that while 0 % and 3 % (women and men respectively) expressed underestimates and 20 % and 23 % were accurate in their perceptions of the norm, 80 % and 73 % overestimated the prevalence of heavy weekend drinking among peers. Also in this study, women were three times as likely, and men were more than twice as likely, to overestimate the 3 month prevalence of alcohol-induced vomiting among peers compared to underestimating its prevalence. In reports of the number of days drinking per month, students attending a university in Scotland estimated that their peers drank more than twice as often as indicated by self reports (McAlaney & McMahon, 2007). Likewise, students' average perception of the frequency of other students being drunk each month was double that reported by students at this university. Similarly, a study of 11 institutions across seven provinces of Canada revealed that regardless of the actual drinking norm at each school, students tended to misperceive the norm in each context with 84 % overestimating the frequency of consumption and 76 % overestimating the amount consumed at parties and bars (Perkins, 2007). Arbour-Nicitopoulos, Kwan, Lowe, Taman, and Faulkner (2010) reported a perception vs. actual norm gap for tobacco and marijuana as well as alcohol in research among Canadian university students at one university. Data collected on university students in five Latin American countries (Brazil, Chile, Colombia, Honduras, and Peru) revealed overestimations of the prevalence of using tobacco, marijuana, and cocaine, and although the prevalence of alcohol use was not typically overestimated, drinking was perceived to be much more frequent than the actual frequency norm (Bustamante et al., 2009).

Although the research on misperceived substance use norms is most prevalent for college student samples, the phenomenon is not characteristic of higher education populations alone. A state-wide study of 21–34-year-olds (only a small portion of them were current students) in Montana found massive overestimates of peer drinking and driving behaviours (Perkins, Linkenbach, Lewis, & Neighbors, 2010). Extensive misperception of exaggerated peer norms for alcohol, tobacco, and other drug use has also been documented in secondary schools with students ranging in age from 10 to 18 based on diverse samples collected in the USA (Beck & Treiman, 1996; Haines, Barker, & Rice, 2003; Linkenbach & Perkins, 2003; Perkins & Craig, 2003a), in four countries (Hungary, Slovakia, Czech Republic, and Romania) of

Eastern Europe (Page, Ihasz, Hantiu, Simonek, & Klarova, 2008; Page, Ihasz, Simonek, Klarova, & Hantiu, 2006), and in Tasmania (Hughes, Julian, Richman, Mason, & Long, 2008).

Following upon the documentation of overestimation of peer support for and use of alcohol, tobacco, and illicit drugs, other research on adolescents and young adults has directed the study of misperceived norms to other areas of health-related problem behaviours. For example, a study in eight secondary schools in the western USA revealed that students overestimated the norm for the amount of sugar-sweetened beverages consumed by other students in their class year for each class year cohort in each school (J. Perkins, Perkins, & Craig, 2010a). A study of secondary students in a large London, England borough revealed substantial misperception of peer body weight norms where 34 % of males and 32 % of females substantially overestimated the same gender and class year weight norm and 37 % of males and 43 % of females underestimated the peer norm (J. Perkins, Perkins, & Craig, 2010b). Multiple studies of students attending universities located in diverse regions of the USA have documented misperception of norms regarding sexual activity (Lewis, Lee, Patrick, & Fossos, 2007; Lynch, Mowrey, Nesbitt, & O'Neill, 2004; Martens et al., 2006; Scholly, Katz, Gascoigne, & Holck, 2005; Seal & Agostinelli, 1996). These studies document students substantially overestimating the frequency of various peer sexual behaviours such as vaginal and anal intercourse and oral sex, overestimating peers' number of sexual partners within the last year, and underestimating the prevalence of peer protective behaviours such as condom use. Other studies have uncovered misperceptions of peer norms (overestimates) concerning male perpetration of intimate partner violence among male perpetrators of such violence (Neighbors, Walker, et al., 2010), and among male college students, misperceptions of peer norms (underestimates) of both males' and females' beliefs about the importance of consent in sexual activity and willingness to intervene against sexual violence (Fabiano, Perkins, Berkowitz, Linkenbach, & Stark, 2003). Similarly, overestimates of peer attitudes tolerating bullying, overestimates of peer perpetration of bullying, and underestimates of the willingness of peers to report bullying to teachers or authorities were found in each of five middle schools studied in the state of New Jersey in the USA (Perkins, Craig, & Perkins, 2011).

Perceived Norms and Personal Behaviour

Even though misperceptions of norms were pervasive, some individuals perceived peer norms with a good deal of accuracy in the research described above, and among those who did not, there was considerable variation in the degree of misperception in many instances. Thus, we must also consider the implications of this variation in perceived peer norms. What is the potential effect of differing perceptions of the norm among individuals who all share the same peer group? If norms do exert a force on individual behaviour, and if the classic sociological dictum holds true that situations or circumstances perceived as real are real in their consequences

(Thomas & Thomas, 1928), then it is reasonable to expect that this variation in perceived norms (or the degree of accuracy in estimating the norm) will be significantly associated with variation in personal behaviour within the group. That is, at least part of the impact of social norms is likely to occur through one's impression of the norm regardless of one's accuracy in estimating its objective existence. Perceptions of the norm, be they accurate or inaccurate, must be taken as important in their own right since people act on their perceptions in addition to acting within an objective normative world. Thus, if misperceptions are pervasive and if perceived norms are influential, the result may be a classic "reign of error" (Merton, 1957) where a false definition of the situation evokes new behaviour as misperceptions control personal action in various populations and contexts.

An association between the perceived norm and personal behaviour is, indeed, commonly demonstrated in empirical research on adolescent/young adult health and problem or risk-related behaviours. For example, several studies using data collected in a variety of secondary schools and colleges in different countries demonstrate a significant positive association between the variation in what students believe to be the norm among other students at their school regarding alcohol use and variation in personal drinking behaviour (cf. Clapp & McDonnell, 2000; Hansen, 1993; Hughes et al., 2008; McAlaney & McMahon, 2007; Neighbors, Lee, Lewis, Fossos, & Larimer, 2007; Page et al., 2008). One nationwide study of 140 colleges and universities throughout the USA with a sample of 17,562 students (Perkins & Wechsler, 1996) found that the perception of more permissive peer attitudes (injunctive norm) was significantly associated with greater personal negative consequences of alcohol use after controlling for the student's personal attitude regarding alcohol consumption and variation in alcohol abuse among schools in the study. Research in diverse settings has also demonstrated a significant positive correlation between perceived peer norms and other personal behaviours including: (a) tobacco use among students attending a French university (Franca, Dautzenberg, Falissard, & Reynaud, 2009) and high school students in Eastern European countries (Page et al., 2006), (b) marijuana use among university students at three schools in the northwestern region of the USA (Kilmer et al., 2006), (c) sugar-sweetened beverage consumption in eight secondary schools in the western USA (J. Perkins et al., 2010a), (d) sexual activity and risk-related behaviour in two studies of university students attending schools in different regions of the USA (Lewis et al., 2007; Martens et al., 2006), (e) extent of intimate partner violence among male perpetrators studied in one region of the USA (Neighbors, Walker, et al., 2010), and bullying attitudes and behaviours among middle school students (class years 6–9) in one school in Portugal (Almeida, Correia, & Marinho, 2010) and five schools in an east coast state of the USA (Perkins et al., 2011).

Five additional studies demonstrating an association between perceived peer norms and personal risk or problem behaviour among youth and young adults are especially important to single out here as they examined the degree of association between the actual local peer norm and personal behaviour simultaneously with the degree of association between the perceived peer norm and personal behaviour. This type of multivariate analysis requires a large data base with data collected

from several sites providing variation in actual norms along with the variation in perceived norms that commonly occurs. Perkins et al. (2005) provide such an assessment with data collected from more than 72,000 students attending 130 colleges and universities in the USA. Based on the aggregate personal behaviours of students at each school, the actual norm for amount that students drink in social situations at each school was used to predict personal quantities consumed while the student's perceptions of the peer norm at his or her school simultaneously was introduced as a predictor of personal consumption in a multivariate analysis. Student perception of the local campus drinking norm was the strongest predictor of the amount of alcohol personally consumed in comparison with the effects of the actual campus drinking norm and all other demographic variables included in the study. A subsequent study of more than 5,000 university students attending 11 institutions across Canada (Perkins, 2007) produced a parallel result with perception of the peer drinking norm at the local institution providing the strongest predictor of personal consumption among all variables and a much larger association than that of the actual norm with personal consumption. Another study focused on alcohol consumption specifically among 4,258 college student-athletes in 15 colleges and universities located across the USA and analyzed the predicted effects of both male and female actual and perceived norms (Perkins & Craig, 2012). Perception of the male student-athlete drinking norm was the strongest predictor of personal drinking levels for both genders in comparison with the effects of the actual male and female norm and demographic variables. The perceived female student-athlete drinking norm was also a strong predictor of female but not male consumption. A fourth study examined sugar-sweetened beverage consumption (SSBC) in a sample of 3,831 secondary school students representing 29 grade level cohorts from grades 6 to 12 in eight schools in the western USA (J. Perkins et al., 2010a). Here, again the perceived norm for SSBC was by far the strongest predictor of personal SSBC compared to all socio-demographic variables included in the study, and the estimated actual SSBC norm for the students' local grade cohort had no significant effect. The perceived norm independently accounted for 34 % of the explained variation in personal SSBC while all other variables accounted for only 5 % of the personal SSBC variation. The fifth study examined the association of secondary school students' personal body mass index (BMI) with the estimated actual and perceived average weights of the same-sex students in one's class year in one's local school (J. Perkins et al., 2010b). The data from 2,104 students represent 37 same gender and class year cohorts drawn from 14 secondary schools in a large and ethnically diverse borough of London, England. For males, personal BMI was significantly predicted simultaneously by both their perceptions of the peer (same gender and class year) norm and by actual cohort norms with about equal predictive power. For females, personal BMI was significantly and strongly predicted by perceived same gender and class year norms while actual norms were insignificant in predicting BMI.

The strong empirical association between perceived peer norms and personal behaviour, as found in the many cross-sectional studies described above, does not confirm causality of course. It is quite reasonable to assume, based on theory, that there may be causal effects in each direction. Just as perceived norms may be partial

determinants of individual behaviour, it is plausible that the individual's personal behaviour may have some determining effect on his or her perceptions of what is the typical behaviour of others. Thus, more complex studies are needed to test the directionality and degree of effect in each direction. One type of analysis investigating this question involves longitudinal data using a cross lagged method of multivariate statistical analysis. In these studies data collected on both the perceived norm and personal behaviour at time 1 are used to simultaneously predict the perceived norm and also personal behaviour at time 2. Using this method the effect of the prior perceived norm, independent of the effect of the prior personal behaviour, can be isolated when predicting later personal behaviour and perceptions of the norm. Thus, the simultaneous potential influences of the perceived norm and personal behaviour on subsequent personal behaviour and the perceived norm can be separated.

Only four studies were found using some type of cross lagged analysis to address this question of the causal direction in the relationship of perceived norms and personal behaviour in the research literature. The results provide varied evidence on how strongly perceived norms determine personal behaviour when controlling for effects in the opposite direction. In a study of college student drinking in one university in the USA, Neighbors, Dillard, Lewis, Bergstrom, and Neil (2006) found support for a mutual influence model but also found stronger support for personal conformity to perceived peer norms in contrast with the process of personal behaviour shaping perceptions. In another longitudinal study of university student drinking (Cullum, Armeli, & Tennen, 2010) that collected data over three time points, the structural equation analysis also found results supporting each directional pathway. In this study the effect of perceived norms on personal consumption was consistent across multiple time points, but more limited in the size of the effect at each time in comparison with the effects of personal behaviour on perceptions. Another longitudinal study of college student drinking (Pedersen, LaBrie, & Hummer, 2009) examined pre-abroad factors that predicted drinking behaviour while studying abroad. Both pre-abroad intentions of drinking (personal attitude) and pre-abroad perceptions of study-abroad drinking (perceived norms of future peer environment) were associated with subsequent drinking abroad. However, pre-abroad perceptions predicted actual study-abroad drinking over and above one's intentions. Furthermore, only study participants with higher pre-abroad perceived norms of abroad drinking significantly increased their drinking while abroad, thus providing additional support for perception's impact on personal behaviour. Juvonen, Martino, Ellickson, and Longshore (2007) used 7th grade perceived norms and personal behaviour to predict personal alcohol and marijuana use among students in the 8th grade in 21 schools in the state of South Dakota in the USA. In this study, students' previously perceived peer norms significantly predicted personal alcohol use but not marijuana use. When students' 7th grade recall of the number of times peers had offered them alcohol in their lifetime and how often they were around peers who drank alcohol (what might be interpreted as related to perceptions of more proximal peer norms), the effect size of the perceived 7th grade norm on personal 8th grade drinking was diminished and statistical significance was lost.

Other tests for the causal impact of perceived norms on personal behaviour that provide substantial supporting evidence are found in the studies using some form of experimental longitudinal design. The intervention or experimental condition is some type of experimenter action to change perceptions of the norm followed by the examination of subsequent changes in personal behaviour. Results of these studies are reviewed in the subsequent section of this chapter when considering how misperceived norms may be changed.

The Dynamic View of Perceived Norms

Although the pervasiveness of misperceived norms and its potential detrimental effects on the well-being of youth and young adults has been established, the review of these findings, as introduced thus far, is not intended to convey a static image of norms or perceptions of norms and their associations with personal behaviour. Misperceptions of norms do emerge for individuals and may change, which, in turn, may bring changes in individual action. Thus, it is important to consider the dynamics that produce the misperceptions, the potential for altering misperceptions, and the effects that may result from such changes.

Causes of Misperceived Norms

A multiplicity of causes has been cited for the explanation of misperceived norms. Psychologists often rely on the concepts of "pluralistic ignorance" and "false consensus" to explain the discrepancy between actual and perceived norms for youth risk behaviour (cf. Berkowitz, 2005; Prentice & Miller, 1993; Schroeder & Prentice, 1998). Simply put, pluralistic ignorance posits a psychological tendency among many people to think of themselves as somewhat different from most others, and thus the potential for an overall discrepancy between the aggregate of personal attitudes and behaviours and what is perceived as average or most typical of others. Furthermore, if the majority believe themselves to be in the minority, they will then tend to keep their opinions private and restrict their actual behaviour preferences when acting publicly—a process that makes actual norms less visible, further exacerbating misperceptions and further restricting the revelation of real personal preferences for behaviour in a pernicious manor. They may not only participate in the misperceived norm occasionally to publicly disguise their opposition, but also participate in the encouragement and enforcement of others' participation as a means of further (and more convincingly) communicating to peers their apparent, albeit insincere, allegiance (Willer, Kuwabara, & Macy, 2009). False consensus posits a process whereby individuals exhibiting minority attitudes and behaviours tend to think that most others are like themselves. This process is predicted from a

psychological viewpoint as a "self-serving bias", a way to reinforce their own views and actions, and also from a social psychological viewpoint as the result of "selective exposure" to a greater prevalence of deviant behaviour in one's immediate environment or personal relationships.

Relying solely on the combination of pluralistic ignorance and false consensus to explain the phenomenon of misperceived norms for youth risk behaviour is problematic, however, for several reasons. First, there is no prior predictive explanation of who is likely to be a victim of pluralistic ignorance, or a victim of false consensus if motivated by a "self-serving bias." Rather, these are theorized conditions for misperceiving the norm often attached to individuals as a label once we know whether their own personal attitudes or behaviours reflect the actual norm or reflect a non-normative position. Second, these theoretical constructs do not account for patterns of misperception such as that reported about frequency and quantity of alcohol use among university students where individuals with personal consumption levels substantially below the normative behaviour still tend to overestimate (rather than underestimate) the norm (even though they do not typically overestimate it as much as those who are above the norm in personal consumption). Third, the concepts of pluralistic ignorance and false consensus do not directly address from a sociological vantage point how institutional and cultural products also contribute to these misperceived norms.

I have argued in detail elsewhere for another set of concepts providing a theoretical model (Perkins, 1997, 2002, 2003a) of misperceived norms in the research on health and well-being among youth and young adults. The model incorporates both psychological and sociological phenomena that in combination theoretically explain the emergence and persistence of misperceived norms. The model, very briefly described here, posits three levels of processes that create and mutually reinforce misperceptions. The first level based on cognition processes looks to the psychological tendency to mistakenly assume that extreme behaviour, when occasionally or even rarely observed in others we do not know well, reflects their dispositions and common ways of behaving. These psychological "attribution errors" are made when only incomplete or superficial information about peers is available. They become more substantial as the distance between the perceiver and those being observed is greater because the perceiver does not have the opportunity to observe others who are not intimate contacts in a variety of contexts, where such observations might otherwise moderate their impressions of what is typical of others. This phenomenon is secondly coupled with the tendency of people to remember vivid and extreme behaviour (such as the risk and problem behaviours discussed in this chapter) more often than normative behaviour and then to talk about it disproportionately in social conversation. (Consider the hundreds of words and expressions used in various youth and adult cultures to describe inebriation in comparison to the very few words available to describe the condition of sobriety even though sobriety is normative in virtually all youth and adult populations including university students in the vast majority of social circumstances). Thus, the social psychology of conversation patterns brings disproportionate attention to these non-normative attitudes and behaviours amplifying the sense that they are

pervasive, while talk about what is actually most common gets little attention. Finally, a third level of distortion is introduced through cultural communications. Many forms of television, film, and website entertainment accentuate risk behaviours as attractive and commonplace. Likewise, news media concentrates on drawing public attention to (and sensationalizing) the high-risk and problem behaviours within a population (as the media slogan goes, "if it bleeds, it leads"). Thus, exposure to disproportionate media content of youth risk behaviours can create the impression that these behaviours are much more commonplace than is the reality as popular culture focuses almost entirely on images and stories of the unusual and extreme behaviours, both locally and in the larger society. Taken together, distortion in perceptions of the norm produced by psychological tendencies and social conversation patterns are reinforced by the socio-cultural level of human experience and vice versa.

The theoretical causes of misperceived norms discussed above suggest that the creation and reinforcement of misperceptions is a perpetual process in most instances. If, among youth for example, (1) there is the tendency to erroneously attribute risk behaviours, when occasionally observed, to typical dispositions or inclinations of peers, (2) social conversation amplifies one's sense of the prevalence of the behaviour, and (3) the cultural media simultaneously hype its prevalence, then the predicted result would be increasing misperception of the norm in the direction of the problem behaviour. Simultaneously, if misperceptions of the norm do contribute to the encouragement and growth of attitudes and behaviours that are misperceived to be normative, then one should logically predict a steady increase in the problem behaviour until it becomes the actual norm or perhaps until it becomes virtually universal. And yet as one might rightly point out, problem rates among youth overall do not inevitably increase over time, possibly leading one to the impression that the suggested process of an at least partially self-fulfilling prophecy is not taking place. In fact, however, the dynamic growth (or perverse increase) in the problem behaviour in the wake of widespread and growing misperceptions is indeed taking place during the adolescent years, but youth do not stay in the same constant and isolated group through time. That is, we rarely watch one age group of peers monitoring both their perceptions of the norm and their personal behaviours over a lengthy period of time. But we do see steady increases in perceived norms and personal behaviours regarding the prevalence of alcohol and drug use across school years as adolescents move into older grades. So at any one moment, if we examine an entire school or a particular year level (grade), the norms and exaggerated perceptions of norms may appear to be fairly constant when compared to a previous assessment of the school or same year level (grade). But beneath the surface (or from a longitudinal point of view) the picture is different. Overtime, more individuals in a year level (grade) cohort may initiate a behaviour in response to their perceptions of what is normative as they prepare to move (anticipatory socialization), and then do move, into the next levels. Thus, more of them will begin to adopt the perceived normative behaviour thinking they need to do so to "fit in" at the next level. The process does not continue indefinitely to a point where everyone is really doing it because students move beyond the peer intensive school environments to

new normative groups in the proverbial "real" world of occupations, military service or newly emerging families with more diverse reference groups and where their perceptions of what is normative (be they correct or incorrect) are altered.

Interventions

Just as there is a dynamic nature to the creation, growth and impact of misperceived norms as they evolve through time in the adolescent's and the young adult's life experience, there also exists the possibility of change in perception and behaviour due to interventions designed to alter perceptions of the norm. The "social norms approach" (Perkins, 2003b) to health promotion has been introduced in a variety of contexts as a positive implementation of social norms theory to reduce problem behaviour based on the principle that much of the problem behaviour is encouraged and perpetuated by pervasive misperception that the problem behaviour is the norm. Thus, a successful intervention to reduce or correct misperceptions of the norms should have the reverse effect (reducing problems) as some people begin to shift their attitudes and behaviours in accordance with their new (more accurate) perceptions of the norm. More individuals may be willing to behave in accordance with their underlying attitudes if they come to believe that the majority of peers support them and they may be more willing to voice their opinions or intervene as well, providing a further counter to the remaining misperceptions of the norms and problem behaviour among peers. Those who previously may have flagrantly exhibited extreme problem behaviour believing their actions were widely valued may be less likely to do so or do so publicly, thereby assisting in the further reduction of the problematic misperceived norm.

Interventions employing this strategy use a variety of techniques in attempts to correct misperceptions, typically based on previously gathered credible information about actual norms or based on techniques that expose the actual norms of a group in the course of the intervention. These techniques commonly include the use of print and electronic media to advertise actual norms, the implementation of group workshops, orientation programs, or online interactive programs providing presentations of findings on actual norms or interactive exercises to reveal the actual dominant attitudes and behaviours of the peer group.

Experimental evidence supporting this theory and practical approach to achieve change has grown substantially in the last two decades as applied to a variety of issues involving the promotion of health and well-being in schools and communities. The most extensive supporting evidence comes from interventions designed to reduce misperceptions of high-risk drinking as the norm among university students in the USA. Several studies have used a pre/post quasi-experimental design to assess perceived norms, the frequency and quantity of personal alcohol consumption, or the experience of alcohol-related negative consequences at one or more time points prior to and again after an intervention. The first of these studies was conducted at a mid-sized university in the Midwestern region (Haines & Spear, 1996).

Initially, data collected at two time points (from one academic year to the next) while not conducting a social norms intervention showed no significant change in alcohol measures (perceptions of heavy drinking as the norm and personal heavy drinking rates). In the next year an intervention to reduce misperceptions of the norm was introduced with a widespread print media campaign about accurate norms and student staged theatrics to further publicize the correct data about local norms. The prevalence of misperception that heavy drinking was the norm immediately dropped significantly from 69 to 57 % as did the prevalence of personal heavy drinking from 45 to 38 % (a statistically significant rate of change decrease of 16 %). The study reported continued declines over the following 2 years of intervention resulting in a 24 % decline in the heavy drinking measure (rate of change) after 3 years of intervention while the national prevalence of heavy drinking among college students remained unchanged. The intervention at this school to reduce misperceptions and the assessments were subsequently continued for a total of 9 years following the baseline assessment (Haines & Barker, 2003) ending with an overall drop in the misperceived heavy drinking norm from 69 to 33 % cutting misperceptions by more than half (−52 % rate of change) and a reduction in personal heavy drinking from 45 to 25 % (−44 % rate of change).

Other colleges and universities conducted experimental interventions and assessments using similarly intense print media campaigns and supplementing them with electronic media and other communication strategies to communicate actual norms over the next several years with similar results. For example, assessments after 3 and 5 years of intervention at a small private liberal arts college in the Northeast saw continuing declines resulting in a 32 % overall reduction (rate of change) in heavy drinking (Perkins & Craig, 2002, 2003b). A large public university in the Southwest experienced a 29 % decrease (rate of change) in heavy drinking in a 3-year pre/post assessment (Johannessen & Glider, 2003). A midsized university in the Northwest observed a statistically significant 21 % reduction (rate of change) in heavy drinking in the year following its social norms intervention and after an assessment showing no change in heavy drinking rates over the previous pre-intervention 5-year time period (Fabiano, 2003). A midsized university in the mid-Atlantic eastern region experienced yearly declines in the prevalence of heavy drinking resulting in a 25 % reduction (rate of change) 3 years after the pre-intervention baseline measure (Jeffrey et al., 2003). These schools also reported significant reductions on several measures of perceived norms and other measures of problem drinking and negative consequences in these studies.

More recently a study of the impact of a social norms intervention at a midsized Southeastern university has demonstrated that as the project expanded its communication strategy about accurate norms throughout the university's student body over a 6-year period, yearly declines in negative consequences of drinking followed (Turner, Perkins, & Bauerle, 2008). In 2001, 44 % of students experienced multiple negative consequences, but by 2006 the rate had dropped to 25 %. One large study of 18 schools throughout the USA was able to construct an experiment with random assignment of half of the schools as control sites for comparison. After 3 years the social norm intervention sites revealed relatively lower

perceptions of drinking norms and lower rates of personal problem drinking compared to the control schools, a finding that did not exist at the start of the experiment (DeJong et al., 2006).

In addition, several social norms intervention programs have successfully targeted specific sub-populations of students by communicating actual norms of the group (e.g. first-year students, residence hall residents, fraternity and sorority members, and student-athletes) within the university environment through media campaigns (Berkley-Patton, Prosser, McCluskey-Fawcett, & Towns, 2003; Mattern & Neighbors, 2004), peer-based programming efforts (Cimini, Page, & Trujillo, 2002), group feedback using wireless keypads (LaBrie, Hummer, Grant, & Lac, 2010), computer-delivered normative feedback (Lewis & Neighbors, 2007; Neighbors, Larimer, & Lewis, 2004), workshop or counseling formats to reduce misperceptions and problem drinking (Barnett, Far, & Mauss, 1996; Borsari & Carey, 2000; Steffian, 1999) or a combination of these strategies (Perkins & Craig, 2006). Successful intervention experiments are also reported with students identified as heavy drinkers and students mandated for programs due to alcohol policy violations (Agostinelli, Brown, & Miller, 1995; Collins, Carey, & Sliwinsky, 2002; Cunningham, Wild, Bondy, & Lin, 2001; Doumas, McKinley, & Book, 2009; Neighbors et al., 2004) as well as with students living in small residential groupings (Schroeder & Prentice, 1998).

Certainly many of the intervention studies described above have some methodological limitations such as the lack of a randomized control group for comparison over time as used in classical experimental designs. Also, many studies are based on research conducted in single institutional contexts, thereby limiting the strength and generalizability of findings. The similar pattern of positive results found, however, in so many studies conducted at diverse sites over time gives much credence to the argument that interventions to change perceived norms can, in turn, change behaviour. Still it must be noted that, although accumulated intervention studies present a very large body of supporting evidence for the malleability and influence of perceived norms, not all social norms interventions to reduce high-risk drinking among college students have been successful in demonstrating support for the approach. Most of the unsuccessful interventions, however, used weak or problematic communications strategies or short time frames that did not produce a reduction in the level of misperceptions of the norm (Granfield, 2002; Thombs, Dotterer, Olds, Sharp, & Raub, 2004; Werch et al., 2000), a result that social norms theory posits should yield no change in the personal drinking levels (Perkins, 1997). Thus, reports of failed experiments do not typically present results countering the fundamental theoretical assumptions of the social norms model (Thombs et al., 2004) (i.e. that a correction or change in normative perception affects personal behaviour). Rather, they most often reflect problems of (1) very low intervention dosage (i.e. limited exposure to social norm messages due to insufficient intervention intensity or duration), (2) lack of credible data for messages, (3) an overly narrow focus on a target group without reducing misperceptions of the broad student population (Perkins, 2003c), or (4) confusing presentations regarding actual norms (Russell, Clapp, & DeJong, 2005). One report described as a "failed" study (Clapp, Lange, Russell, Shillington, & Voas, 2003)

actually found results of significantly lowered misperception in a student residence hall when results were compared to another residence hall with no intervention that was used a the control group, but the study did not find a significant reduction in actual drinking levels. However, the intervention was done only inside the residence hall and with only one simple print message, and then impact was assessed after only 6 weeks. Thus, obtaining substantial behavioural change might not be realistic, and yet the critical personal behaviour measures all moved in the expected direction compared to the control group, suggesting that some impact may have taken place but not enough to be significant and avoid a possible Type II error (Perkins, 2006).

One study of 14 institutions randomly assigned to a social norms intervention or control school condition (DeJong et al., 2009) reported no difference at the end of the experiment that was attempting to replicate a previous study of 18 randomly assigned schools where an intervention effect had been found (DeJong et al., 2006). One possible explanation reported for the failure to replicate the impact of an intervention communicating accurate norms in the second wave study as compared to the first wave of schools studied was that the second wave of schools were disproportionately institutions where a high density of alcohol outlets existed close to the campus and alcohol consumption was relatively high compared to the first wave of schools studied. Thus, the second study concluded that social norms interventions may not be as effective in environments with a high density of alcohol outlets and the pervasive promotion of alcohol consumption. This result may simply mean, however, that the intensity of exposure to correct normative information may need to be increased in these circumstances beyond what was a minimal intervention dosage. Intervention schools in this study were given just $2,000 for the creation and purchase of media advertisements while some of the institutions had populations of 20,000–40,000 students so the message dosage per student from media was inevitably very limited. Successful school interventions in other studies using mass media marketing would not uncommonly spend at least ten times that amount to gain enough exposure in schools of that size and in much smaller schools.

Finally, we can note other evidence that interventions to change perceptions of norms can bring about corresponding changes in problem drinking and other problem behaviours in school and community settings beyond the university context. An experiment conducted throughout the State of Montana in the USA (Perkins, Linkenbach, et al., 2010) assigned a portion of the counties as intervention counties and others as control counties. The study subsequently conducted an intensive mass media campaign communicating the accurate norm in the experimental counties that most (four out of five) young adult (21–34 years old) Montanans do not drink and drive (based on data from statewide surveys) when the misperception was pervasive that most would drink and drive in a typical month. After 18 months misperceptions about the norm were reduced, the willingness to use designated non-drinking drivers increased, and drinking and driving decreased in the intervention counties compared to the control counties. In another experiment middle school students in 12 schools in southern California were assigned to one of four experimental conditions (resistance skill training, normative education to reduce

misperceived peer norms about the prevalence of drug use, a combination of both skill training and normative education, and a control condition with neither type of education) during the school year. As a result, alcohol, cigarette, and marijuana use were reduced due to the effect of normative education with no significant effect of resistance skills training (Hansen, 1993; Hansen & Graham, 1991). A pre/post assessment of tenth grade students exposed to a social norms campaign in two Illinois high schools demonstrated significant reductions in alcohol use and tobacco use over a 2-year time period (Haines et al., 2003). In an 8 month media campaign throughout selected counties in the state of Montana teenagers were targeted with the message that "7 out of 10 are tobacco free" and related normative messages that most teenagers do not use tobacco. Experimental counties at the end of the trial showed an initiation rate for tobacco use of only 10 % among teens not previously using tobacco compared with a 17 % rate in control counties that did not receive the normative messages (Linkenbach & Perkins, 2003). In a social norms intervention at five middle schools in New Jersey addressing misperceptions about the prevalence of peer bullying attitudes and behaviour and willingness to report bullying to teachers, the campaigns were effective in reducing erroneous perceptions and changing attitudes and behaviours in a more positive direction (Perkins et al., 2011). Among the five sites, the schools where greater campaign exposure was reported were also the schools where, over time, greater increases in accurate perceptions of norms and greater decreases in personal perpetration and support for bullying occurred.

Finally, it should be noted that some evidence, albeit much more limited, also exists beyond the field of youth risk behaviour prevention supporting social norms theory's prediction that interventions communicating actual norms will bring change. For example, experiments in adult populations have demonstrated that conveying information about descriptive and injunctive norms can impact environmental concerns such as littering, recycling, energy consumption, and protection of environmental resources (Cialdini et al., 1990; Nolan, 2011; Schultz, 1999; Schultz, Khazian, & Zaleski, 2008).

Current and Future Issues for the Study of Perceived Norm Dynamics

Although there is much accumulated evidence supporting the claims that misperceptions of norms regarding risk behaviours are pervasive and can be altered, in turn, producing change in individual behaviour, several important theoretical issues remain where empirical investigation is quite limited. Space constraints for this chapter will only permit a brief description of these areas in need of further investigation.

One important question involves the comparison of proximal and distal reference group norms. It is not uncommon for theory and empirical research to point out that proximal norms (e.g. norms of one's more immediate friendship network) are more

influential than distal norms (e.g. norms of one's entire school population) (cf. Cho, 2006; Thombs, Ray-Tomasek, Osborn, & Olds, 2005). Presumably, people pay greater attention to and are more directly influenced by the norms of a close group of peers that they care about more strongly and interact with more intensely. Multivariate analyses sometimes show that when friend norms (actual or perceived) are entered along with norms of peers in general (actual or perceived) to simultaneously predict personal behaviour, the norms of close friends account for most or almost all of the explained variation in personal behaviour (Maddock & Glanz, 2005). Some studies have shown that young people can also misperceive the norms of their close friends leading to some speculation that addressing those misperceptions may be more effective in producing change. But such a decision is not that straightforward. First, it must be acknowledged that identifying friend norms and then communicating these back to the individual is a much more complex endeavour when large populations are involved and this usually requires the loss of anonymity in survey research which may be problematic regarding sensitive issues. Second, the extent of misperception of close friend norms will not be as large as the gap observed between actual and perceived norms of peers in general in the local population. This is because the psychological process of making attribution errors leads to greater error and exaggeration about people who are in more distal groupings (Perkins, 1997). Therefore, while the influence of close peer norms may be greater, the extent of misperception, and thus the possible extent of change (correction) in the perceived norm will likely be less. Even though the distal peer norm may be less influential, there is likely to be massive misperception allowing more potential change to occur in the perceived norm. So addressing both proximal and distal misperceptions hold some promise for change in individuals' behaviour (LaBrie, Hummer, Neighbors, & Larimer, 2010; Larimer et al., 2009; Neighbors et al., 2008). Future research also needs to consider how the misperception of each type of norm may contribute to or reinforce the misperception of the other norm. Furthermore, future research needs to examine the potential interactive effects of misperceived norms at both levels, and thus the possible additional effect of addressing both misperceptions simultaneously in interventions.

A second related line of needed inquiry involves questions about the effect of social network density and group identification and how these factors might mediate the effect of misperceived norms. It is theoretically plausible that even among groups representing the same social sphere—for example, all other students in one's classroom in a secondary school—a more tight knit or interconnected network among students in the class may produce greater conformity to the perceived norm, and thus possibly greater change, if misperceptions are reduced.

A third possibility involves the study of variation in individual attitudes and dispositions concerning the importance of peers. Various psychological and sociocultural characteristics may lead individuals to be more or less group oriented in terms of relying on the group for personal guidance. Thus, correcting misperceptions by providing feedback about accurate group norms may be more or less influential on the individuals depending on their personal propensity or desire to conform to the group (Neighbors, LaBrie, et al., 2010).

Finally, research has begun to explore gender dynamics in understanding misperceived norms and their influence on the individual. For example, some theoretical speculation and limited research among adolescents and young adults has suggested that same gender norms might be a more powerful influence depending on the topic (Korcuska & Thombs, 2003; Lewis et al., 2007; Lewis & Neighbors, 2004, 2007). Still other work suggests that in cultural circumstances where male attitudes and behaviours are valued more highly in general, that perhaps perceptions of the male norm may be more highly associated with what is perceived as the non-gender specific norm and more influential on personal behaviour for both genders (Lewis & Neighbors, 2006; Pedersen & LaBrie, 2008; Perkins & Craig, 2012).

To conclude, these emerging areas of research provide many directions for future inquiry as to how misperceptions of norms develop, become solidified as the perceived reality, affect subsequent behaviour, and can be changed through interventions to alter perceptions producing subsequent change in behaviours. Conducting such research in diverse cultural contexts and on risk behaviours beyond alcohol and substance abuse also provide a wide terrain for new exploration of the "reign of error" and how to confront it in promoting human well-being.

References

Agostinelli, G., Brown, J. M., & Miller, W. R. (1995). Effects of normative feedback on consumption among heavy drinking college students. *Journal of Alcohol and Drug Education, 25*(1), 31–40.

Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology, 52*, 27–58.

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology, 32*(47), 665–683.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Upper Saddle River, NJ: Prentice-Hall.

Ajzen, I., & Madden, T. (1986). Prediction of goal-directed behavior: Attitudes, intentions and perceived behavioral control. *Journal of Experimental Social Psychology, 22*, 453–474.

Almeida, A., Correia, I., & Marinho, S. (2010). Moral disengagement, normative beliefs of peer group, and attitudes regarding roles in bullying. *Journal of School Violence, 9*, 23–36.

Arbour-Nicitopoulos, K. P., Kwan, M. Y. W., Lowe, D., Taman, S., & Faulkner, G. E. J. (2010). Social norms of alcohol, smoking, and marijuana use with a Canadian university setting. *Journal of American College Health, 59*(3), 191–196.

Asch, S. E. (1951). Effects of group pressure on the modification and distortion of judgements. In H. Guetzkow (Ed.), *Groups, leadership and men*. Pittsburgh, PA: Carnegie Press.

Asch, S. E. (1952). *Social psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs, 70*(9), 1–70.

Baer, J. S., & Carney, M. M. (1993). Biases in the perceptions of the consequences of alcohol use among college students. *Journal of Studies on Alcohol and Drugs, 54*, 54–60.

Baer, J. S., Stacy, A., & Larimer, M. (1991). Biases in the perception of drinking norms among college students. *Journal of Studies on Alcohol, 52*, 580–586.

Barnett, L., Far, J., & Mauss, A. (1996). Changing perceptions of peer norms as a drinking reduction program for college students. *Journal of Alcohol and Drug Education, 41*, 39–62.

Beck, K. H., & Treiman, K. A. (1996). The relationship of social context of drinking, perceived social norms, and parental influence to various drinking patterns of adolescents. *Addictive Behaviors, 21*, 633–644.

Berkley-Patton, J., Prosser, E., McCluskey-Fawcett, K., & Towns, C. (2003). A social norms approach to reducing alcohol consumption among college freshman. *NASPA Journal, 40*(4), 24–37.

Berkowitz, A. D. (2005). An overview of the social norms approach. In L. C. Lederman & L. P. Stewart (Eds.), *Changing the culture of college drinking: A socially situated health communication campaign* (pp. 193–214). Cresskill, NJ: Hampton Press.

Borsari, B., & Carey, K. B. (2000). Effects of a brief intervention with college student drinkers. *Journal of Consulting and Clinical Psychology, 68*, 728–733.

Borsari, B., & Carey, K. B. (2001). Peer influences on college drinking: A review of the research. *Journal of Substance Abuse, 13*, 391–424.

Burrell, L. F. (1990). College students' recommendations to combat abusive drinking habits. *Journal of College Student Development, 31*, 562–563.

Bustamante, I. V., Carvalho, A. M. P., Oliveira, E. B., Olieira, H. P. J., Figueroa, S. D. S., Vasquez, E. M. M., et al. (2009). University students' perceived norms of peers and drug use: A multicentric study in five Latin American countries. *Review of Latin American Enfermagem, 17*, 838–843.

Carey, K. B., Borsari, B., Carey, M., & Maisto, S. (2006). Patterns and importance of self-other differences in college drinking norms. *Psychology of Addictive Behaviors, 20*(4), 385–393.

Cho, H. (2006). Influences of norm proximity and norm types on binge drinkers: Examining the under-examined aspects of social norms interventions on college campuses. *Journal of Substance Use, 11*(6), 417–429.

Cialdini, R. B., Reno, R. R., & Kallgren, C. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*(6), 1015–1026.

Cimini, M., Page, J. C., & Trujillo, D. (2002). Using peer theatre to deliver social norms information: The Middle Earth Players program. *The Report on Social Norms, 2*(1), 1–7.

Clapp, J. D., Lange, J. E., Russell, C., Shillington, A., & Voas, R. (2003). A failed social norms marketing campaign. *Journal of Studies on Alcohol, 64*, 409–414.

Clapp, J. D., & McDonnell, A. L. (2000). The relationship of perceptions of alcohol promotion and peer drinking norms to alcohol problems reported by college students. *Journal of College Student Development, 41*, 19–26.

Collins, S., Carey, K., & Sliwinsky, M. (2002). Mailed personalized normative feedback as a brief intervention for at-risk college drinkers. *Journal of Studies on Alcohol, 63*, 559–567.

Cullum, J., Armeli, S., & Tennen, H. (2010). Drinking norm–behavior association over time using retrospective and daily measures. *Journal of Studies on Alcohol and Drugs, 71*, 769–777.

Cunningham, J. A., Wild, T. C., Bondy, S. J., & Lin, E. (2001). Impact of normative feedback on problem drinkers: A small-area population study. *Journal of Studies on Alcohol, 62*, 228–233.

DeJong, W., Schneider, S. K., Towvim, L. G., Murphy, M., Doerr, E., Simonsen, N., et al. (2006). A multisite randomized trial of social norms marketing campaigns to reduce college drinking. *Journal of Studies on Alcohol, 67*(6), 868–879.

DeJong, W., Schneider, S. K., Towvim, L. G., Murphy, M., Doerr, E., Simonsen, N., et al. (2009). A multisite randomized trial of social norms marketing campaigns to reduce college student drinking: A replication failure. *Substance Abuse, 30*(2), 127–140.

Doumas, D., McKinley, L., & Book, P. (2009). Evaluation of two web-based alcohol interventions for mandated college students. *Journal of Substance Abuse Treatment, 36*(1), 65–74.

Fabiano, P. M. (2003). Applying the social norms model to universal and indicated alcohol interventions at Western Washington University. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse* (pp. 83–99). San Francisco, CA: Jossey-Bass.

Fabiano, P. M., Perkins, H. W., Berkowitz, A., Linkenbach, J., & Stark, C. (2003). Engaging men as social justice allies in ending violence against women: Evidence for a social norms approach. *Journal of American College Health, 52*(3/4), 105–112.

Foss, R., Marchetti, L., & Holladay, K. (2001). *Development and evaluation of a comprehensive program to reduce drinking and impaired driving among college students, Report No. DOT HS 809 396*. Washington, DC: National Highway Traffic Safety Administration, US Dept. of Transportation.

Franca, L. R., Dautzenberg, B., Falissard, B., & Reynaud, M. (2009). Are social norms associated with smoking in French university students? A survey report on smoking correlates. *Substance Abuse Treatment, Prevention, and Policy, 4*, 4.

Granfield, R. (2002). Can you believe it? Assessing the credibility of a social norms campaign. *The Report on Social Norms* (Working Paper #2).

Haines, M., & Barker, G. P. (2003). The Northern Illinois University experiment: A case study of the social norms approach. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse* (pp. 21–34). San Francisco, CA: Jossey-Bass.

Haines, M., Barker, G. P., & Rice, R. (2003). Using social norms to reduce alcohol and tobacco use in two midwestern high schools. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse* (pp. 235–244). San Francisco, CA: Jossey-Bass.

Haines, M., & Spear, S. (1996). Changing the perception of the norm: A strategy to decrease binge drinking among college students. *Journal of American College Health, 45*, 134–140.

Hancock, L., & Henry, N. (2003). Perceptions, norms, tobacco use in college residence hall freshman: Evaluation of a social norms marketing intervention. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse* (pp. 135–153). San Francisco, CA: Jossey-Bass.

Hansen, W. B. (1993). School-based alcohol prevention programs. *Alcohol Health and Research World, 17*(1), 54–60.

Hansen, W. B., & Graham, J. W. (1991). Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing conservative norms. *Preventive Medicine, 20*, 414–430.

Hechter, M., & Opp, K.-D. (2001). What have we learned about the emergence of social norms? In M. Hechter & K.-D. Opp (Eds.), *Social norms* (pp. 394–415). New York, NY: Russell Sage.

Horne, C. (2001). Sociological perspectives on the emergence of social norms. In M. Hechter & K.-D. Opp (Eds.), *Social norms* (pp. 3–33). New York, NY: Russell Sage.

Hughes, C., Julian, R., Richman, M., Mason, R., & Long, G. (2008). Harnessing the power of perception. *Youth Studies Australia, 27*(2), 26–35.

Jeffrey, L. R., Negro, P., Miller, D., & Frisone, J. D. (2003). The Rowan University social norms project. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse* (pp. 100–110). San Francisco, CA: Jossey-Bass.

Johannessen, K., & Glider, P. (2003). The University of Arizona's campus health social norms media campaign. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse* (pp. 65–82). San Francisco, CA: Jossey-Bass.

Juvonen, J., Martino, S., Ellickson, P. & Longshore, D. (2007). "But others do it!" Do misperceptions of schoolmate alcohol and marijuana use predict subsequent drug use among young adolescents? *Journal of Applied Psychology, 37*(4), 740–758.

Kilmer, J., Walker, D., Lee, C., Palmer, R., Mallett, K., Fabiano, P. M., et al. (2006). Misperceptions of college student marijuana use: Implications for prevention. *Journal of Studies on Alcohol, 67*(3), 277–281.

Korcuska, J., & Thombs, D. (2003). Gender role conflict and sex-specific drinking norms: Relationships to alcohol use in undergraduate women and men. *Journal of College Student Development, 44*(2), 204–216.

Kypri, K., & Langley, J. D. (2003). Perceived social norms and their relation to university student drinking. *Journal of Studies on Alcohol, 64*(6), 829–834.

LaBrie, J., Hummer, J., Grant, S., & Lac, A. (2010). Immediate reductions in misperceived social norms among high-risk college student groups. *Addictive Behaviors, 35*, 1094–1101.

LaBrie, J., Hummer, J., Lac, A., & Lee, C. (2010). Direct and indirect effects of injunctive norms on marijuana use: The role of reference groups. *Journal of Studies on Alcohol and Drugs, 71*, 904–908.

LaBrie, J., Hummer, J., Neighbors, C., & Larimer, M. (2010). Whose opinion matters? The relationship between injunctive norms and alcohol consequences in college students. *Addictive Behaviors, 35*, 343–349.

Larimer, M., Kaysen, D., Lee, C., Kilmer, J., Lewis, M., Dillworth, T., et al. (2009). Evaluating level of specificity of normative referents in relation to personal drinking behavior. *Journal of Studies on Alcohol and Drugs*, (Suppl. 16), 115–121.

Lewis, M., Lee, C., Patrick, M., & Fossos, N. (2007). Gender-specific normative misperceptions of risky sexual behavior and alcohol-related risky sexual-behavior. *Sex Roles, 57*, 81–90.

Lewis, M., & Neighbors, C. (2004). Gender-specific misperceptions of college student drinking norms. *Psychology of Addictive Behaviors, 18*(4), 334–339.

Lewis, M., & Neighbors, C. (2006). Who is the typical college student? Implications for personalized normative feedback intervention. *Addictive Behaviors, 31*(11), 2120–2126.

Lewis, M., & Neighbors, C. (2007). Optimizing personalized normative feedback: The use of gender-specific referents. *Journal of Studies on Alcohol and Drugs, 68*, 228–237.

Linkenbach, J., & Perkins, H. W. (2003). Most of us are tobacco free: An eight-month social norms campaign reducing youth initiation of smoking in Montana. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse* (pp. 224–234). San Francisco, CA: Jossey-Bass.

Lynch, J., Mowrey, R., Nesbitt, G., & O'Neill, D. (2004). Risky business: Misperceived norms of sexual behavior among college students. *NASPA Journal, 42*(1), 21–35.

Maddock, J., & Glanz, K. (2005). The relationship of proximal normative beliefs and global subjective norms to college students' alcohol consumption. *Addictive Behaviors, 30*, 315–323.

Martens, M., Page, J., Mowry, E., Damann, K., Taylor, K., & Cimini, M. D. (2006). Differences between actual and perceived student norms: An examination of alcohol use, drug use, and sexual behavior. *Journal of American College Health, 54*(5), 295–300.

Mattern, J., & Neighbors, C. (2004). Social norms campaigns: Examining the relationship between changes in perceived norms and changes in drinking levels. *Journal of Studies on Alcohol, 65*, 489–493.

McAlaney, J., Bewick, B., & Hughes, C. (2010). The international development of the 'Social Norms' approach to drug education and prevention. *Drugs: Education, Prevention and Policy, 18*(2), 81–89.

McAlaney, J., & McMahon, J. (2007). Normative beliefs, misperceptions, and heavy episodic drinking in a British student sample. *Journal of Studies on Alcohol and Drugs, 68*(3), 385–392.

McCabe, S. E. (2008). Misperceptions of non-medical prescription drug use: A web survey of college students. *Addictive Behaviors, 33*, 713–724.

Merton, R. K. (1957). The self-fulfilling prophecy. In R. K. Merton (Ed.), *Social theory and social structure* (pp. 421–436). New York, NY: Free Press.

Neighbors, C., Dillard, A., Lewis, M., Bergstrom, R., & Neil, T. (2006). Normative misperceptions and temporal precedence of perceived norms and drinking. *Journal of Studies on Alcohol, 67*(3), 290–299.

Neighbors, C., LaBrie, J. W., Hummer, J. F., Lewis, M. A., Lee, C. M., Desai, S., et al. (2010). Group identification as a moderator of the relationship between perceived social norms and alcohol consumption. *Psychology of Addictive Behaviors, 241*(3/4), 522–528.

Neighbors, C., Larimer, M., & Lewis, M. (2004). Targeting misperceptions of descriptive drinking norms: Efficacy of a computer-delivered personalized normative feedback intervention. *Journal of Consulting and Clinical Psychology, 72*(3), 434–447.

Neighbors, C., Lee, C., Lewis, M., Fossos, N., & Larimer, M. (2007). Are social norms the best predictor of outcomes among heavy-drinking college students? *Journal of Studies on Alcohol and Drugs, 68*(4), 556–565.

Neighbors, C., O'Connor, R., Lewis, M., Chawla, N., Lee, C., & Fossos, N. (2008). The relative impact of injunctive norms on college student drinking: The role of reference group. *Psychology of Addictive Behaviors, 22*(4), 576–581.

Neighbors, C., Walker, D., Mbilinyi, L., O'Rourke, A., Edleson, J., Zegree, J., et al. (2010). Normative misperceptions of abuse among perpetrators of intimate partner violence. *Violence Against Women, 16*(4), 370–386.

Nolan, J. M. (2011). The cognitive ripple of social norms communications. *Group Processes & Intergroup Relations, 14*(5), 689–702.

Page, R. M., Ihasz, F., Hantiu, J., Simonek, J., & Klarova, R. (2008). Social normative perceptions of alcohol use and episodic heavy drinking among central and eastern European adolescents. *Substance Use and Misuse, 43*, 361–373.

Page, R. M., Ihasz, F., Simonek, J., Klarova, R., & Hantiu, I. (2006). Cigarette smoking, friendship factors, and social norm perceptions among central and eastern European high school students. *Journal of Drug Education, 36*(3), 213–231.

Page, R. M., Scanlan, A., & Gilbert, L. (1999). Relationship of the estimation of binge drinking among college students and personal participation in binge drinking: Implications for health education and promotion. *Health Education, 30*, 98–103.

Pedersen, E., & LaBrie, J. (2008). Normative misperceptions of drinking among college students: A look at the specific contexts of prepartying and drinking games. *Journal of Studies on Alcohol and Drugs, 69*(3), 406–411.

Pedersen, E., LaBrie, J., & Hummer, J. (2009). Perceived behavioral alcohol norms predict drinking for college students while studying abroad. *Journal of Studies on Alcohol and Drugs, 70*(6), 924–928.

Perkins, H. W. (1994). The contextual effect of secular norms on religiosity as moderator of student alcohol and other drug use. In M. L. Lynn & D. O. Moberg (Eds.), *Research in the social scientific study of religion* (pp. 187–208). Greenwich, CT: JAI Press.

Perkins, H. W. (1997). College student misperceptions of alcohol and other drug norms among peers: Exploring causes, consequences, and implications for prevention programs. In *Designing alcohol and other drug prevention programs in higher education: Bringing theory into practice* (pp. 177–206). Newton, MA: The Higher Education Center for Alcohol and Other Drug Prevention, U.S. Department of Education.

Perkins, H. W. (2002). Social norms and the prevention of alcohol misuse in collegiate contexts. *Journal of Studies on Alcohol*, (Suppl. 14), 164–172.

Perkins, H. W. (2003a). The emergence and evolution of the social norms approach to substance abuse prevention. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse* (pp. 3–18). San Francisco, CA: Jossey-Bass.

Perkins, H. W. (2003b). *The social norms approach to preventing school and college age substance abuse*. San Francisco, CA: Jossey-Bass.

Perkins, H. W. (2003c). The promise and challenge of future work on the social norms model. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse* (pp. 280–296). San Francisco, CA: Jossey-Bass.

Perkins, H. W. (2006). Success and failure in social norms interventions (author response to Clapp and Lange). *Journal of Studies on Alcohol, 67*(3), 483–484.

Perkins, H. W. (2007). Misperceptions of peer drinking norms in Canada: Another look at the "reign of error" and its consequences among college students. *Addictive Behaviors, 32*, 2645–2656.

Perkins, H. W., & Berkowitz, A. D. (1986). Perceiving the community norms of alcohol use among students: Some research implications for campus alcohol education programming. *International Journal of the Addictions, 21*(9 & 10), 961–976.

Perkins, H. W., & Craig, D. W. (2002). *A multifaceted social norms approach to reduce high-risk drinking*. Newton, MA: The Higher Education Center for Alcohol and Drug Prevention.

Perkins, H. W., & Craig, D. W. (2003a). The imaginary lives of peers: Patterns of substance use and misperceptions of norms among secondary school students. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse* (pp. 209–223). San Francisco, CA: Jossey-Bass.

Perkins, H. W., & Craig, D. W. (2003b). The Hobart and William Smith Colleges experiment: A synergistic social norms approaching print, electronic media and curriculum infusion to reduce collegiate problem drinking. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse* (pp. 35–64). San Francisco, CA: Jossey-Bass.

Perkins, H. W., & Craig, D. W. (2006). A successful social norms campaign to reduce alcohol misuse among college student-athletes. *Journal of Studies on Alcohol, 67*(6), 880–889.

Perkins, H. W., & Craig, D. W. (2012). Student-athletes' misperceptions of male and female peer drinking norms: A multi-site investigation of the "reign of error". *Journal of College Student Development, 53*(3), 367–382.

Perkins, H. W., Craig, D. W., & Perkins, J. M. (2011). Using social norms to reduce bullying: A research intervention among adolescents in five middle schools. *Group Processes and Intergroup Relations, 14*(5), 703–722.

Perkins, H. W., Haines, M., & Rice, R. (2005). Misperceiving the college drinking norm and related problems: A Nationwide study of exposure to prevention information, perceived norms and student alcohol misuse. *Journal of Studies on Alcohol, 66*, 470–478.

Perkins, H. W., Linkenbach, J., Lewis, M., & Neighbors, C. (2010). Effectiveness of social norms media marketing in reducing drinking and driving: A statewide campaign. *Addictive Behaviors, 35*, 866–874.

Perkins, H. W., Meilman, P. W., Leichliter, J. S., Cashin, J. R., & Presley, C. A. (1999). Misperceptions of the norms for the frequency of alcohol and other drug use on college campuses. *Journal of American College Health, 47*(6), 253–258.

Perkins, J., Perkins, H. W., & Craig, D. W. (2010a). Misperceptions of peer norms as a risk factor for sugar-sweetened beverage consumption among secondary school students. *Journal of the American Dietetic Association, 110*, 1916–1921.

Perkins, J., Perkins, H. W., & Craig, D. W. (2010b). Peer weight norm misperception as a risk factor for being over and underweight among UK secondary school students. *European Journal of Clinical Nutrition, 64*, 965–971.

Perkins, H. W., & Wechsler, H. (1996). Variation in perceived college drinking norms and its impact on alcohol abuse: A nationwide study. *Journal of Drug Issues, 26*(4), 961–974.

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology, 64*, 243–256.

Russell, C., Clapp, J. D., & DeJong, W. (2005). Done 4: Analysis of a failed social norms marketing campaign. *Health Communication, 17*(1), 57–65.

Scholly, K., Katz, A., Gascoigne, J., & Holck, P. (2005). Using social norms theory to explain perceptions and sexual health behaviors of undergraduate college students: An exploratory study. *Journal of American College Health, 53*(4), 159–166.

Schroeder, C. M., & Prentice, D. A. (1998). Exposing pluralistic ignorance to reduce alcohol use among college students. *Journal of Applied Social Psychology, 28*, 2150–2180.

Schultz, P. W. (1999). Changing behavior with normative feedback interventions: A field experiment on curbside recycling. *Basic and Applied Social Psychology, 21*(1), 25–36.

Schultz, P. W., Khazian, A., & Zaleski, A. (2008). Using normative social influence to promote conservation among hotel guests. *Social Influence, 3*(1), 4–23.

Seal, D. W., & Agostinelli, G. (1996). College students' perceptions of the prevalence of risky sexual behaviour. *AIDS Care, 8*(4), 453–466.

Sherif, M. (1936). *The psychology of social norms*. New York, NY: Harper.

Sherif, M. (1972). Experiments on norm formation. In E. P. Hollander & R. G. Hunt (Eds.), *Classic contributions to social psychology*. New York, NY: Oxford University Press.

Steffian, G. (1999). Correction of normative misperceptions: An alcohol abuse prevention program. *Journal of Drug Education, 29*, 115–138.

Thomas, W., & Thomas, D. (1928). *The child in America*. New York, NY: Alfred Knopf.

Thombs, D. L., Dotterer, S., Olds, R. S., Sharp, K., & Raub, C. G. (2004). A close look at why one social norms campaign did not reduce student drinking. *Journal of American College Health, 53*(2), 61–68.

Thombs, D. L., Olds, R. S., & Snyder, B. M. (2003). Field assessment of BAC data to study late-night college drinking. *Journal of Studies on Alcohol, 64*, 322–330.

Thombs, D., Ray-Tomasek, J., Osborn, C., & Olds, R. S. (2005). The role of sex-specific normative beliefs in undergraduate alcohol use. *American Journal of Health Behavior, 29*, 342–351.

Turner, J., Perkins, H. W., & Bauerle, J. (2008). Declining negative consequences related to alcohol misuse among students exposed to a social norms marketing intervention on a college campus. *Journal of American College Health, 57*(1), 85–93.

Werch, C. E., Pappas, D. M., Carlson, J., DiClemente, C., Chally, P., & Sinder, J. (2000). Results of a social norms intervention to prevent binge drinking among first-year residential college students. *Journal of American College Health, 49*, 85–92.

Willer, R., Kuwabara, K., & Macy, M. W. (2009). The false enforcement of unpopular norms. *American Journal of Sociology, 115*(2), 451–490.

Chapter 3
Norms and Beliefs: How Change Occurs

Cristina Bicchieri and Hugo Mercier

Societies are rife with negative, damaging practices, from open defecation to female genital cutting (FGC), endemic in many developing countries, to corruption and violence against women and children that we also witness in many Western societies. The theoretical and practical challenge we face is twofold. On the one hand, we want to explain what generates and supports such practices. On the other, we want to find ways to change them permanently. We will argue here that social norms play an important role in both tasks. Often norms support or embed certain practices, so that eliminating the latter involves changing the former. Sometimes, however, norms have to be created in order to eliminate a negative practice and support a new one, as we know of several widely practiced behaviors that are not supported by norms, but can be changed by introducing them. To understand what we mean by "practice" and "norm," we shall next refer to Bicchieri (2006) definition of social norms, a definition that allows to shed light on the way norms are supported, and on ways we may act to change them.

Social Norms

There are many behavioral regularities we engage in, from brushing teeth in the morning to adopting dress codes, from staying in line to buy a movie ticket to observing rules of fairness in allocating PhD slots. Some such regularities are behaviors that we adopt and keep following irrespective of what others do, or expect us to do. I brush my teeth every morning because I believe in certain hygiene principles, and the fact that most of the American population does the

C. Bicchieri (✉) • H. Mercier
University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: cb36@sas.upenn.edu

M. Xenitidou and B. Edmonds (eds.), *The Complexity of Social Norms*,
Computational Social Sciences, DOI 10.1007/978-3-319-05308-0_3,

same has no impact on my decision. I care about germs and bacteria, not about what others do or don't do. When I go to a party, however, I usually care about the local dress code, may ask about it, and try to conform to what I expect others to wear. Dressing differently would not be a tragedy, just a cause for embarrassment, even if it is obvious that the other guests are tolerant and would not judge me negatively. In this case, what I expect others to wear has an influence on my decision about clothes. When I am in line to buy a movie ticket, I do not try to cut it or jump ahead. I expect everyone to patiently wait his turn, and I know that I am expected to behave accordingly. There is a sense that everyone *ought to* behave in an appropriate way, and we all get mad if someone tries to cut the line and jump ahead. Expecting this generalized reproach is enough to keep us all obeying the rule.

All of the above examples describe widely adopted behavioral regularities, the difference among them laying in the reasons why we follow them. In the case of dressing codes and staying in line, our expectations about what others do or will do are paramount in giving us a reason to behave in that particular way. Yet there is a difference between a simple empirical expectation (all will wear a black tie, all wait in line) and a normative expectation (all those who wait in line believe I ought to wait patiently in my place). In the first case, expecting a certain behavior gives me a definite reason to follow it; in the latter case, I need a further inducement in the form of a sanction (negative, in this example) to decide that it is better not to cut the line. Social norms, it has been argued (Bicchieri, 2006), are behavioral rules supported by a combination of empirical and normative expectations. Individuals have a *conditional* preference for obeying social norms, provided they hold the right expectations[1].

Empirical expectations are always important, since in their absence we may be tempted to disobey social norms, especially those that demand behavior that may conflict with self-interest. Norms of cooperation, reciprocity and fairness, for example, may lose their grip when we are faced with widespread transgressions. In that case, the force of the norm is greatly diminished. Yet, even when widely followed, social norms may require, to be obeyed, the further belief that others think we ought to obey them, and may be prepared to punish our transgression. Such *normative expectations* always accompany social norms and are usually consistent with our empirical expectations of widespread compliance.

As we shall see, conceiving of norms as supported by, and in a sense constituted of, individuals' expectations offers many theoretical advantages. For one, we now have an operational definition of "social norm" that allows us to make predictions and to experimentally test whether a change in expectations results in a change in behavior. We can also assess the presence of social norms by asking people about their second-order beliefs about what others think the appropriate behavior is, and check for the mutual consistency of these beliefs (Bicchieri & Chavez, 2010). We can, and this is the topic of this paper, devise specific interventions to effect norm

[1] Conditional preferences distinguish social from moral or religious norms, where one would choose to conform irrespective of what others are expected to do, or think one ought to do.

change by acting upon the expectations that support the norm we wish to eradicate or, when it is a new norm we want to establish, work at creating new expectations, and focus on those factors that will bridge expectations and behavior. If indeed expectations, both empirical and normative, are crucial to the existence and stability of a norm, it follows that a change in expectations will always induce a change in compliance and, when the change in expectations is widespread, the abandonment of a norm. For those interested in the removal of a negative norm, or the establishment of a new, positive norm, the issue of collective belief change thus becomes of paramount importance.

Changing Empirical Expectations: The Pitfalls

How easy is it to change people's empirical expectations? First of all, individuals should *observe* or at least reasonably *expect* different behavior in a large enough number of relevant people (i.e., people whose behavior and judgment they care about). Notice that there are many cases in which such observation/expectations would prove difficult to come by. Take for example norms about private behavior, such as sexual mores. In this case, we may have widespread, private disagreement with the standing norms, and a significant amount of secret deviance (Schank, 1932). Yet, because public deviance may be costly, we would observe public, open allegiance and support for the norms in question. These cases are typical of *pluralistic ignorance*, a cognitive state in which one believes one's attitudes and preferences are different from those of similarly situated others, even if public behavior is identical (Allport, 1924; Miller & McFarland, 1991). In all these cases, individuals engage in social comparison with others who are similarly situated. Others' behavior is observable, or at least the consequences of behavior are observable, in that if there are few or no pregnancies out of wedlock one would be justified in assuming that sex outside marriage is uncommon and condemned. In all these cases, transparent communication is impossible, as the social situation is one in which the norms in question are thought to be widely adopted and strongly endorsed, and hence, the fear of embarrassment and ostracism that would follow an open declaration of disagreement keeps people in line. Typically people assume that others' behavior is consistent with their attitudes and preferences, and therefore, from observing widespread compliance each will infer that everybody else endorses the social norm, which in turn can only reinforce public allegiance to it.

Such cases of pluralistic ignorance are quite common, even when behavior is public (as opposed to private), such as Prohibition support (Robinson, 1932), the "conservative lag" in behavior toward integration (O'Gorman, 1975), or a "liberal leap" such as the sexual revolution in the 1960s (Klassen, Williams, & Levitt, 1989). Studies of gang members (Matza, 1964), prison guards (Klofas & Toch, 1982), and prison inmates (Benaquisto & Freed, 1996), as well as school teachers (Packard & Willower, 1972) show that the social norms about proper behavior that are widely shared by all these communities are often regarded by their very members as too

strict or even plainly wrong, but nobody dares to question the shared rules for fear of negative sanctions. It has been shown (Bicchieri & Fukui, 1999) that it may take a small number of "trendsetters" who question the standing norm and start behaving differently to effect a major change. But this would mean that we have to move our explanation a step up, in that we need to explain how change in behavior for the trendsetters came about.

Another possibility is change that comes, so to speak, from above. Imagine the case of a government injunction: From now on, FGC is abolished. We have plenty of experience, especially with developing countries, that such injunctions rarely work. It is interesting to note that, on the contrary, an injunction to shift driving to the right side of the road would (and has been) completely successful. Why? A widely announced change in traffic rules is expected to be followed by all drivers. It is in the interest of each individual driver to coordinate with others and knowing that, one can trust that other drivers will comply with the injunction to drive, say, on the right side of the road. This case is one of a shift in conventions. As discussed elsewhere (Bicchieri, 2006), conventions are quite different from social norms, in that they are supported by empirical expectations of compliance, and a preference to follow the convention provided one expects most (or all) others to comply with it. Thus, a government diktat would work for conventions, but be more problematic in case of social norms.

To move away from a shared norm, we need the assurance that we will not suffer negative consequences. This is because social norms are also supported by normative expectations, i.e., the expectation that others believe we *ought* to behave in a given way, and may sanction us (in a negative way) if we stray. Changing norms thus presents us with a collective action problem, as nobody wants to incur the negative sanctions involved in a transgression. *Prima facie*, it would appear that external interventions, in the form of government interventions, may *facilitate* behavioral changes, by taking away the stigma connected with disobeying a widely held social norm. For example, if FGC is widely practiced in a community, then being the first to abandon it would expose the family to significant damages. For one, the uncut girl would not find a husband, and would become the target of negative perceptions[2]. The entire family would suffer negative consequences, as a family that does not cut its girls would be seen as openly flaunting shared norms, and would thus be ostracized.

It would thus seem that introducing laws that prohibit that practice, and thus establish new sanctions, would be a successful measure, as it would alter the cost and benefit of the targeted behavior by changing expectations and the perceptions of what incurs disapproval, and even change a person's own preferences and create guilt, especially when there is a shared norm of obeying the law. Public opposition to the existing norms would become less costly, and therefore, we should see the target behavior eventually disappear. This view embodies the traditional economic analysis of law, an analysis that focuses on its role in changing the cost and benefit

[2] The Saleema case study in Somalia points to the fact that the only word traditionally used for the uncut girl was "ghalfa," which roughly means prostitute (Hadi, 2006).

of targeted behavior: people are predicted to abide by the law if sanctions are sufficiently severe and tend to break the law if sanctions for doing so are too mild. Yet this view is too simplistic, in that it assumes a host of conditions that need to be present in order for the legal solution to be effective. The question whether laws bring about social change hinges on factors such as legitimacy, procedural fairness, and how the law is originated and enforced.

People who view the law as legitimate are more likely to comply with it even though this contradicts their interests. A legitimate law is not just one that ensues from a legitimate, recognized authority. It must also be the case that the procedures through which authorities make decisions are seen as fair, that the law is consistently enforced, and that the enforcers are perceived as honest. So for example the sporadic campaigns that are launched to enforce the laws during politically sensitive periods, such as in pre-electoral times, are not taken too seriously, and the corruption of local enforcers is a powerful delegitimizing influence. Furthermore, individuals' opportunity to take part in the decision-making process, present their arguments, being listened, and having their views considered by the authorities would seem to offer an especially strong incentive to abide by the law.

Legitimacy thus results in respect for the authorities, and a sense of obligation to obey them. Yet, even assuming that the authority that enacts and enforces the law is perceived as legitimate, perhaps the most important factor that determines successful enforcement is a shared sense that the existing legal arrangements are *as they ought to be*, in that they do not appear so distant from existing social norms as to lose credibility.

> If the law strays too far from the norms, the public will not respect the law, and hence will not stigmatize those who violate it. Loss of stigma means loss of the most important deterrent the criminal justice system has. (Stuntz, 2000)

In other words, the law should approximate popular views, otherwise the threat to seek enforcement will not be credible. Platteau (2000) and Aldashev, Imane, Platteau, and Wahhaj (2010a, 2010b) give a series of examples of laws that were successful precisely because they were sufficiently close to shared social norms: in Gabon and Senegal, instead of banning polygamy, the initial marriage contract allowed the choice of monogamy or polygamy. In Ghana, to protect women and children's inheritance rights, a moderate law proved more effective than previous extreme law. In Bogota, where high firearm mortality was common, Mockus, the major of Bogota, decided to ban guns on weekends only, sending a strong signal but also realistically understanding that a moderate legal injunction would be easier to enforce and obey. Dan Kahan (2000) discussion of "gentle nudges" and "hard shoves" similarly points out that if a new legal norm imposes harsh penalties against a widely accepted social norm, police will become less likely to enforce the law, prosecutors will be less likely to charge and juries to convict, with the effect of reinforcing the existing norm that we wanted to change. Milder penalties are much more effective, and enforceable, thus leading to a progressive condemnation and abandonment of the "sticky norm."

In sum, the legal approach to norm change can help change empirical expectations, but only under rather strict conditions. Individuals will abandon a shared

social norm only if they believe that others are changing, too. This belief must be accompanied by a credible change in sanctions, in that the original negative social sanction for *not* following the norm will be substituted by a new, credible negative sanction for following it. In this case, normative expectations would change, too.

Deliberation

A stepping-stone in the process of norm change is affecting people's empirical expectations. If someone believes that others will act in a certain way with regard to a given norm—follow it, say—that person is likely to follow it herself, irrespective of whether she thinks this is the best thing to do otherwise (Bicchieri & Muldoon, 2010). Prima facie, it may thus seem that what really matters is behavior: what people do, not what they say ought to be done. After all, as economists are wont to point out, talk is cheap (Farrell & Rabin, 1996). Yet numerous experiments have demonstrated the power of discussion to promote pro-social behavior by focusing participants on "good" norms (see, for review, Balliet, 2010; Sally, 1995).

One of the contexts in which norms play an important role is solving commons dilemmas. In an idealized, laboratory version of a commons dilemma, participants are given some endowment money and a choice to either keep that money to themselves or invest it in a common pool. All the money invested in the common pool is then multiplied by some amount (larger than 1) and then equally redistributed across all participants. Overall profits are maximized when everybody contributes their whole endowment. Yet each individual is better off letting all the others contribute while keeping her endowment to herself. As a result, when the game is played in repeated rounds, contributions to the common pool rapidly decrease to a negligible level, when participants are not allowed to talk to each other, that is. If the participants can communicate prior to making their decisions, the level of cooperation can remain very high for as many rounds as the experimenter is willing to go (Ostrom, Walker, & Gardner, 1992). When participants can communicate, they are able to focus on (and follow) a norm of contribution to the common pool.

What happens in these discussing groups that make participants more likely to cooperate? Part of it is the result of low-level factors; simply interacting with other people from the group makes participants more likely to cooperate with them, even if that interaction is as minimal as looking each other in the eyes (Kurzban, 2001). But several experiments have demonstrated that the bulk of the effect comes from the ability to make promises (e.g., Bicchieri, 2002; Orbell, Van de Kragt, & Dawes, 1988). During the discussion, people promise to contribute a given amount and the evidence suggests that the majority of the participants are true to their word. In terms of norms, the effect of promises can be described as a change in empirical and normative expectations. Participants now expect others to behave in a way consistent with their pledges, and they expect that people who renege on their promise will be negatively judged.

A limitation of these experiments is that participants do not have to fight an ingrained norm that would hinder the acceptance of a norm of cooperation. What if there is a preexisting norm that dictates non- or low-cooperation? This is not as farfetched as it may seem. In some cultures, people who contribute too much to a common pool are seen as exerting an undue pressure on others to match their level of contribution and are punished for it (Herrmann, Thöni, & Gächter, 2008). In these circumstances, promises may not be sufficient, for they would be less credible. Knowing that sanctions can be incurred if the promises are kept, other participants may not take them as seriously as they would otherwise. As a result, their empirical expectations may remain unchanged, and they would not be inclined to follow a norm of cooperation.

If we look at real-life cases, the risk of empty promises is even more blatant. FGC, mentioned above, provides a good example (LeJeune & Mackie, 2009). It consists in the ablation of parts of the female genitalia (usually the clitoris, sometimes more) and is typically practiced on relatively young girls, certainly before they get married. The practice of FGC is not an isolated cultural norm. It is embedded in a rich network of beliefs—beliefs about the origins of FGC, its religious justifications, its effect on health (or lack thereof) and so on. Many of these beliefs are normative in nature. The virtue of uncut girls, in particular, is often questioned. As a result, people follow the norm not only because of their empirical expectations—they expect others to do the same—but also because of their normative expectations—they expect to suffer from a variety of sanctions if they fail to follow the norm. In such a context, promises are much less likely to result in a switch in empirical expectations, as they are not very credible. The whole network of beliefs surrounding FGC—responsible for the normative expectations—cannot simply vanish; and, as long as it is present, it is going to make norm change extremely difficult. Even if empirical expectations are a crucial element in norm change, changing empirical expectations without first modifying normative expectations is not always possible.

Discussions and deliberations can also play a critical role in changing normative expectations. The simplest type of change that discussions can bring about is lifting pluralistic ignorance. As mentioned above, people can follow a norm because they believe that others would shun them if they didn't, even if this belief is mistaken. If people were only able to candidly share their feelings about the norm, they may just realize that the whole thing is pointless and stop abiding by it. The solution could therefore be purely endogenous. Often things are unlikely to be that simple though—if a friendly chat would have solved the problem, it is likely that the despised norm would have already disappeared. There are several reasons why the relevant exchange does not take place. In contrived laboratory situations—but also in a few real-life cases—communication may simply be impossible. But the most common hindrance to a candid discussion of the norm is the existence of norms that dictate how one should talk about norms. Going back to the FGC example, even if we assumed that a sizeable part of the population was in fact opposed to the norm, these people would have very little chances of expressing such a view. This could be either

because a specific segment of the population—women, often—is not given much of a public voice, or more drastically because the mere mention of FGC would be a very serious normative breach (LeJeune & Mackie, 2009).

When the norm cannot be freely discussed by all the parties involved, trying to force people to talk about it anyway is likely to backfire. The external agent trying to impose such a discussion would likely be perceived very negatively. Even if the discussion were to take place, it could have damaging consequences. If criticisms of the norm are not allowed, a false impression of consensus can strengthen pluralistic ignorance. Following a discussion that all parties believe to have been frank—except for their own contribution—the norm could even acquire more legitimacy. An exogenous element is thus often required to challenge normative beliefs, either to challenge the normative beliefs themselves, or at least to question the normative beliefs that regulate how the targeted normative beliefs are discussed.

The role of the exogenous agent will be, simply put, to make people change their mind about the relevant normative beliefs. One way to do so is to rely on trust and authority. If a religious or secular leader tells people that some of their normative beliefs are mistaken, they may just take her word for it—especially if the leader is respected by everyone in the relevant community. But in many cases it is not possible to merely rely on trust: people have to be *convinced* that they should change their mind. The main tool for conviction is argumentation, and we presently give a brief account of how arguments can change people's beliefs.

Arguments and Belief Change

Beliefs rarely come by as isolated units; they form complex networks with different types of relationships: consequence, explanation, association, etc. It is possible to describe many of these links in terms of coherence: beliefs are more or less coherent, or consistent, with each other (Thagard, 2002). Inconsistencies are typically the occasion for belief change. When inconsistent beliefs are detected, the mind tries to determine which can be most easily rejected in order to reduce the inconsistency (Festinger, 1957). People can stumble upon these inconsistent beliefs on their own, or they can be made to face their inconsistencies by others. This is what arguments do. Arguments take a belief that the listener accepts—the premise—and show her that this belief is inconsistent with the rejection of the argument's conclusion. When a good argument is offered, it is more consistent for the listener to change her mind about the conclusion than to accept the premise while rejecting the conclusion (Mercier & Sperber, 2011).

Arguments can be more or less explicit. In a very explicit argument, the logical relationship is highlighted with logical connectives ("and") or other connectives ("therefore"). That the strength of the argument should be prominently on display is generally a good thing: it makes the argument easier to understand and more persuasive. Yet explicit arguments can also backfire. If the intent of the speaker is ambiguous in the first place, it is more likely to be perceived as manipulative (Kamalski, Lentz,

Sanders, & Zwaan, 2008). Moreover, explicit arguments may appear threatening. The listener may be unable to muster a sound counterargument while still not being persuaded. Such a situation is likely to arise when the issue is heavily emotionally loaded, as in cases of "moral dumbfounding" (Haidt, Bjorklund, & Murphy, 2000). The listener is then likely to feel threatened by the argument, and to have an antagonistic reaction to the speaker who is challenging her beliefs and making her look irrational.

Arguments can also be mostly implicit. Instead of explicitly making the speaker face her inconsistencies, she can be led to realize on her own that some of her beliefs are in fact conflicting with each other. Social norms are steeped in a thick network of beliefs, attitudes and values. Some of them are more central than others, and highlighting conflicts between beliefs (as well as between beliefs, values and attitudes) must be threaded lightly. Tostan is a nongovernmental organization battling FGC in Senegal and other countries, and they rely in large part on this type of implicit arguments (Gillespie & Melching, 2010). They do not bluntly tell people that their beliefs about FGC are inconsistent with, for instance, their desire to have healthy children. Instead, the importance of some values—such as respect for human life—is first highlighted without reference to FGC. People are made to work out, in a process of collective deliberation, the practical consequences of these values. When this approach is coupled with information about FGC—in particular its health consequences—people can work out on their own the problematic aspects of FGC (Diop et al., 2004)

One of the factors that make some beliefs—such as beliefs about FGC—difficult to change is that they are more "central" than others (see, e.g., Judd & Krosnick, 1982). These beliefs are at the center of a dense network of beliefs, attitudes, and values. Keeping on with the example of FGC, the belief that girls should be cut has explanations and consequences, it may be linked with religious beliefs and social customs, it is embedded in specific rituals, etc. A frontal attack on FGC is unlikely to succeed, as many other beliefs would have to simultaneously evolve. By contrast, more peripheral beliefs are more amenable to arguments. For instance, the belief that FGC is part of the Islamic faith is peripheral both to beliefs about FGC and about Islam (this belief is a rationalization, as Islamic scriptures do not in fact recommend FGC). One of the reasons that tackling a relatively central belief often entails a prolonged process is that many peripheral beliefs have to be modified first. A complementary way to target relatively central beliefs is to use beliefs that are even more central. This is one way of describing a major aspect of Tostan's work with deliberations: trying to show that some central values conflict with the belief that girls should be cut.

Discussions and deliberations often allow people to change their normative beliefs and, therefore, the normative expectations related to an old norm that has to be challenged. Still, even the disappearance of the previous normative expectations may not prove sufficient. There are several reasons this may occur. Agreement that a particular norm is not necessary to fulfill some core beliefs, and indeed may be in conflict with some deeply held values is just a first step, necessary but by no means sufficient, to stably change behavior. People must be convinced that their core

beliefs and values are better served by a new practice. Such new practices may be endorsed by a respected leader, or be the result of extensive group discussion that focuses on alternative solutions. The importance of finding alternatives cannot be overstated. Without the possibility of conceiving viable alternatives, abandoning an established norm is a losing proposition.

In lengthy rounds of collective discussions, people may agree that the old norm should not be upheld, come to envision and agree upon a new practice, and promise to follow it (Haile, 2006). Yet if the consequences of being isolated in keeping one's word are too high, people may be reluctant to do so—especially since they know that others are likely to have the same train of though and therefore to back down as well (a reflection that can be made worse by iteration). In communities that practice FGC, it is virtually impossible for an uncut girl to find a husband. In other words, the costs of not following the old norm are potentially enormous. Even if everybody can be persuaded to promise to forswear the custom, people may still fear that others won't keep their word. What is needed then is the establishment of normative beliefs that will transform the new, agreed upon practice into a social norm.

One way to enact such transformation is to publicly commit to change behavior and promise to move in the newly envisaged direction. Public pledges have many advantages: the promiser is more likely to keep his or her word since not doing it exposes to "loss of face" and possibly also to reputational damage. Knowing the costs of a broken promise makes it credible, and generates the trust necessary to start moving in the new direction. Furthermore, even if some participants are not particularly enthusiastic about the new course, witnessing a large number of people committing to change behavior leads to form new empirical and normative expectations. Public, credible promises have the function of creating a common belief that the new behavior will be implemented, and the *expectation* of such behavior. Creating normative expectations, however, is crucial in establishing the new behavior as a social norm.

In an experiment alluded to earlier, participants were able to reach a high level of cooperation—high contributions to a common pool—simply by discussing the game among themselves and making promises. However, a simple variant of that experiments reveals the limits of simple promises. When the stakes were higher, making defecting more appealing, promises were much less successful at maintaining cooperation (Ostrom et al., 1992). A possible solution is to develop normative expectations by introducing sanctions against defectors. Indeed, participants are willing—eager even—to inflict punishment on defectors, in spite of personal costs. One of the reasons punishments are effective in simple common dilemmas is that their meaning is usually unambiguous. If a participant breaks her word to contribute at least a certain amount, she is likely to understand why she is then punished. Most real-life situations, however, are more intricate, so that one could be punished without knowing exactly why. When this happens, the individual being punished may not know how to improve her behavior or, even if she understands why others think she should be punished, she may be unwilling to change, as she may perceive the punishment as unfair.

Discussions and deliberations can also play a crucial role for the establishment of ways to enforce commitments. When a group of people is trying to institute a norm,

they are likely to realize that some sanctions for norm-breakers are in order. Deliberation is a good way to devise monitoring and punishment devices. If the group members have different incentives and perspectives, their views can be heard and taken into account. The resulting sanctioning scheme will be perceived as more legitimate, and will therefore be more effective. Discussions can also prove critical when the punishment is inflicted, as they facilitate an understanding of why it is inflicted and how it can be avoided in the future (Janssen, Holahan, Lee, & Ostrom, 2010).

Common Knowledge and Tipping Points

For most beliefs the most effective way to change them, and thus eventually change the practices that they support, is through argumentation. For argumentation to be successful, however, two conditions are required. First, the arguer must be able to rest on a set of explicit beliefs and values that is equally well entrenched in the listener and that is inconsistent with the target belief that we want to change. Second, the belief must not be held mainly because other people hold it as well. In this latter case, argumentation is not likely to succeed: as long as one does not see other people from the relevant group changing their mind, one is unlikely to change her beliefs. Social norms, we have argued, are supported by *shared* normative beliefs. Therefore, the process of belief change has to be a *collective* one. People, in other words, have to change their mind together. Group discussion, as opposed to individual discussion, is important because, if a group is confronted with a persuasive argument, and people see others accept it, then they may feel free to accept the argument themselves. Accepting an argument and changing behavior, however, are two different things. One may be convinced by an argument and change one's attitude towards a given norm, but hesitate to change behavior for fear of being in a minority. This means that trying to change the behavior of one person after another is bound to be extremely difficult, if not impossible. For a norm to change, the whole group—or at least a sizeable majority—must be reached.

Deliberation and group diffusion are two complementary and necessary ways to make change happen. Yet there is a tension between the two. We know that deliberation works best in small group settings, but if the relevant group is large, using the "common knowledge of change" approach requires that the entire group changes its mind. In the successful Tostan experience, deliberation in small core groups reaches conclusions that are unstable unless and until the group expects others to follow. Members of the small group have an incentive to recruit more people up to the point at which enough people are ready to adopt a new practice. Typically, the core group organizes diffusion of their discussions into wider arenas. In the African experience so well exemplified by Tostan, diffusion has taken several forms: ordinary discussions with family and friends; meetings with elders, religious leaders, and the women's group; a meeting of the whole community; discussions in nearby communities; and inter-village meetings with delegates from surrounding communities. Spontaneous diffusion, when we let the information circulate of its own accord, often cannot be relied upon until the last phase of the operation.

When a practice is strongly interdependent, often it is not enough for individuals simply to adopt a more favorable *attitude* towards a new alternative. The greater the loss (for example, damage to the daughter's reputation) resulting from a failed effort to shift to a new norm, the more people need to be sure that enough other people in the community will together *act* to adopt the new alternative. All must see that all see that there is change. Since norms are grounded on expectations, what we think others do, and what we think others think we should do, must both change in order for a fracture with the past to occur. We engage in alternative behaviors only if we think other people do so as well, and will judge us well for it. Within a population, it often happens that not everybody follows a norm. When this is the case, people can take the proportion of the population that follows a norm into account in their decisions. Imagine a population in which most people would prefer not to beat their wives, but there's a tradition—a norm in fact—to beat them for even small misdeeds. Furthermore, ideals of masculinity, honor and family values are deeply linked to the practice. At some point, a few individuals may be convinced that beating wives is not the best way to fulfill deeply held values, and they may even decide to abandon the practice. Most others remain unmoved, as the minority is too small. Here core group discussion and organized diffusion would play a crucial role, effecting a gradual change in attitudes. If the minority keeps growing, it may reach a tipping point. At this stage, the minority has grown large enough that most other people feel free to break from the norm and stop beating their wives. Norms often change in this way. Progress is very slow at first, as a few people gradually start to adopt a new norm. But when the tipping point is reached, change can be very sudden. It should then be expected that a slow and steady change in attitudes may not be immediately accompanied by an equally slow and steady change in behavior. On the contrary, behavioral change may be sudden and quite dramatic, and difficult to predict. In the experiences that have accompanied abandonment of FGC, change typically occurs when the population reaches common knowledge that a majority is ready to abandon the old practice. Everybody knows that everybody else knows that the majority of the population is adopting a new practice. There are many ways in which such common knowledge can be achieved: an elaborate public declaration by representatives of interconnected communities; the posting and propagation of a decision by a respected and effective local political authority, or the signing of a flag symbolizing the change by each household in the community. All these are ways to publicly celebrate the change and let everyone know that new expectations are in place.

From the Lab to the Field: Scaling Up Norm Change

In our analysis of norm change and deliberation, we have relied substantially on laboratory experiments, accompanied by real-life examples. Following the lead of scholars such as Elinor Ostrom (1991), we urge for a better integration of fieldwork and laboratory experiments. The results obtained in the laboratory, often with so-called WEIRD participants (participants from Western Educated Industrialized Rich Democratic countries, Henrich, Heine, & Norenzayan, 2010) can not always

be generalized to the field. While it is easy to conjure up examples of disappointing group performance in real life—dreadful committee experiences are burned in our memories—we would like to provide an example in which groups in the field can avoid the pitfalls in which their laboratory counterpart regularly fall.

When psychologists ask participants who agree on some issue to talk about it anyway, the attitudes of the participants tend to polarize. For instance, a group of Republicans talking about a tax increase are likely to pile up arguments against it, arguments that everyone is likely to accept uncritically, providing group members with even more reasons to reject the tax hike. Group polarization is so reliably observed in the lab that Cass Sunstein saw it fit to turn it into the "*law* of group polarization" (Sunstein, 2002). Yet the analysis of real-life cases fails to back up such a strong generalization. Historical analyses of important decisions have shown that groups can start and continue being very cohesive, sharing in the same ideology, without succumbing to groupthink and the polarization that generally ensues (Tetlock, Peterson, McGuire, Chang, & Feld, 1992). Studies of "enclave" deliberation among disempowered groups have shown that despite a lack of heterogeneity, deliberation can allow such groups to find a voice without leading to groupthink or polarization (Karpowitz, Raphael, & Hammond, 2009). A prominent historical example is that of the close-knit group of Quakers who, despite their agreement on the fundamental issue, did not polarize, putting forward instead pragmatic solutions that helped achieve the abolition of slavery in England (see, e.g., Brown, 2006). In-depth studies of such cases are necessary to understand in what respect they differ from the laboratory situations that so reliably produce polarization. One suggestion may be that when the personal stakes of the group members increase, polarization is less likely to ensue. Such an hypothesis would greatly diminish the relevance of the laboratory results obtained so far, but not of experiments in general. Indeed, it would be necessary to test the hypothesis in the laboratory to establish its validity.

As the example of group polarization shows, one must exert caution when extrapolating from the laboratory to the real world. A better interaction and integration of field studies and laboratory experiments will be necessary if we are to reach conclusions that are both sound and relevant.

There is, however, another drawback that the studies we have cited so far share, whether they have been done in the lab or in the field: their relatively small scale. Laboratory experiments only involve a very limited number of participants at a time. The example we have taken of the role of deliberation in norm change—the work of Tostan in fighting FGC—involves significantly larger groups—up to 150 people or more. Yet even with groups of that size, it is relatively easy to imagine that deliberations can affect a substantial part of the group either directly or with at most one level of communication (i.e., someone involved in the debate talking about it with someone who had not taken part in it). Given that norm change requires that a substantial section of the population is ready to change—the famed tipping point—it is not clear exactly how such a process can be scaled up if the goal is to change a norm in an average Western country of several millions inhabitants. This issue is particularly important for policy makers who want to promote behavioral changes in areas such as health or business where entrenched norms stand in the way of progress.

Some processes of norm change are susceptible to scale up relatively easily. Imagine for instance a typical situation of pluralistic ignorance. If people can be made to speak their mind more or less freely when they are surveyed by a pollster, and if the results of the poll can be made public by a trusted source, there is hardly any limit to the size of the population that can be affected. The only constraints are the costs of polling and publicizing the results. In many cases, however, the change has to be deeper than simply making people reveal their true preferences: people have to genuinely change their mind. Deliberation is the best tool to induce belief change, but it doesn't scale up very well: Studies show that as groups grow larger, pre-play communication aimed at inducing cooperative behavior breaks down more easily, as it becomes more difficult to create the trust necessary to support commitments to cooperate. Large groups, it has been argued, could benefit from computer-mediated communication. Yet, even with small numbers, we know that cooperation is more difficult to establish when the means of communication is a computer (Bicchieri & Lev-On, 2007; Bicchieri, Lev-On, & Chavez, 2010). Important aspects of "commitment production," such as coordinating mutual promises, the credibility of promises, and attainment of public knowledge about mutual promising, become problematic in computer-mediated environments. If mutual expectations are crucial in attaining belief (and norm) change, finding the means to achieve a change in expectations should be one of our main goals.

The difficulties we have highlighted mean that deliberation can hardly be the only mean through which a norm can change when large numbers of people are involved. Activists understand this very well, and so they rely on a variety of other media to effect norm change, from ad campaigns to spreading new words that encapsulate a normative statement (such as "homophobic").

Deliberation, however, should not be written off when large numbers are involved. One of the most important movements in recent political science pushes for a more "deliberative democracy" (see, e.g., Elster, 1998; Gutmann & Thompson, 1996). Partisans of deliberative democracy are obviously aware of the scaling up problem, but they can be willing to confront it head on. For instance, Ackerman and Fiskin have suggested that a national "Deliberation Day" should be instituted (Ackerman & Fishkin, 2004). During this day, which would be a national holiday held shortly before an election, all the registered voters would be invited to discuss their views on the upcoming election. While such a project may sound unrealistic now, smaller versions of the same idea have already been implemented. For instance, America*Speaks* organizes debates between small groups of citizens, who then share their results with a larger local group of several hundred people, who then shares these results with other such groups across the country, reaching several thousand people. The goal of such deliberation is not to effect norm change directly, but they can be a crucial step on the way to norm change. People can get a better idea not only of what other people think, but also of why they hold such views (Hansen, 2003). More importantly maybe, people can change their mind about norm-relevant beliefs (e.g., Luskin, Fishkin, & Jowell, 2002).

Aside from these formal debates, deliberation can also play a critical role in norm change through its action in everyday life. Mansbridge has argued that students of deliberative democracy should pay greater attention to the role of "everyday talk" (Mansbridge, 1999). In her study of the feminist movement, she has noted how women have been able to exert an influence on men in their surroundings through ordinary interactions. Such local interactions are influenced by larger trends. For instance, women were able to recruit terms devised by activists—such as "male chauvinist"—in order to make a point quickly and effectively. But, importantly, the multiplication of similar local interactions can also exert a significant effect on the population at large.

Clearly, a lot remains to be done to link the study of local interactions with the application to norm change in large polities. We hope that a better integration of lab experiments and field data, as well as increased dialogue between psychology and the social sciences will help close that gap.

Conclusion

To abandon negative norms, we need to change people's empirical and normative expectations. Discussions and deliberations can be effective means to enact change, as they facilitate the creation of the new empirical and normative expectations that are central to a norm's existence. The positive side-effects of collective deliberations, such as improved interpersonal understanding (Fishkin & Luskin, 2005), increased respect among participants (Gutmann & Thompson, 1996), better solutions to a variety of practical, moral and intellectual problems (Mercier, 2011; Mercier & Landemore, 2012) can prove significant to norm change, but are likely to be peripheral to the main issue. In the present chapter we have confined our analysis to aspects of discussions and deliberations that allow tackling norm change more directly, starting with the ability to change people's empirical expectations. Discussions can change attitudes and clarify what people intend to do. Norm change can sometimes be effected simply by people promising that they will abandon the old norm or follow a new one. For promises to be effective, however, they have to be credible. If people think that others have a strong incentive not to keep their promises, they are unlikely to keep them either. The normative expectations attendant to the old norm can still be in place, making people fear that they would endure sanctions by keeping their word. Through deliberation, an exogenous agent can challenge these normative beliefs, paving the way for an easier transition to a new norm. A new norm can also be favored by the development of normative expectations, potentially accompanied by sanctions for norm violators. Here again, discussions and deliberations should make an important contribution. A punishment scheme that is devised through discussion is perceived as more legitimate, and a punishment accompanied by an explanation is more effective.

References

Ackerman, B., & Fishkin, J. S. (2004). *Deliberation day*. Yale: Yale University Press.

Aldashev, G., Imane, C., Platteau, J.P., & Wahhaj, Z. (2010a). Using the law to change the custom. *Journal of Development Economics*. In press.

Aldashev, G., Platteau, J-P., & Wahhaj, Z. (2010b). Legal reform in the presence of a living custom: An economic approach. *Proceedings of the National Academy of Sciences of the United States.*

Allport, F. (1924). *Social psychology*. Boston: Houghton Mifflin Company.

Balliet, D. (2010). Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution, 54*(1), 39.

Benaquisto, L., & Freed, P. J. (1996). The myth of inmate lawlessness: The perceived contradiction between self and other in inmates' support for criminal justice sanctioning norms. *Law and Society Review, 30*, 481–511.

Bicchieri, C. (2002). Covenants without swords: Group identity, norms, and communication in social dilemmas. *Rationality and Society, 14*(2), 192–228.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. New York: Cambridge University Press.

Bicchieri, C., & Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making, 23*(2), 161–178.

Bicchieri, C., & Fukui, Y. (1999). The great illusion: Ignorance, informational cascades and the persistence of unpopular norms. *Business Ethics Quarterly, 9*, 127–155.

Bicchieri, C., & Lev-On, A. (2007). Computer-mediated communication and cooperation in social dilemmas: An experimental analysis. *Politics, Philosophy and Economics, 6*, 139–168.

Bicchieri, C., Lev-On, A., & Chavez, A. (2010). The medium or the message? Communication richness and relevance in trust games. *Synthese, 176*(1), 125–147.

Bicchieri, C., & Muldoon, R. (2010). Social norms. The Stanford encyclopedia of philosophy. Retrieved Mar, 2011 from http://plato.stanford.edu/entries/social-norms/.

Brown, C. L. (2006). *Moral capital*. Chapel Hill: University of North Caroline Press.

Diop, N., Faye, M. M., Moreau, A., Cabral, J., Benga, H., Cisse, F., et al. (2004). *The Tostan program: Evaluation of a community based education program in Senegal*. Tostan: New York, NY: Tostan – Population Council.

Elster, J. (1998). *Deliberative democracy*. Cambridge: Cambridge University Press.

Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives, 10*, 110–118.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford: Stanford University Press.

Fishkin, J. S., & Luskin, R. C. (2005). Experimenting with a democratic ideal: Deliberative polling and public opinion. *Acta Politica, 40*(3), 284–298.

Gillespie, D., & Melching, M. (2010). The transformative power of democracy and human rights in nonformal education: The case of Tostan. *Adult Education Quarterly, 60*, 477–499.

Gutmann, A., & Thompson, D. F. (1996). *Democracy and disagreement*. Cambridge: Belknap.

Hadi, A. A. (2006). A community of women empowered: The story of Deir El Barsha. In R. M. Abusharaf (Ed.), *Female circumcision* (pp. 104–124). Philadelphia: University of Pennsylvania Press.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished manuscript, University of Virginia.*

Haile, Gabriel D. (2006). A study on community dialogue leading towards abandoning of harmful traditional practices, with special reference to female genital cutting, early marriage, and marriage by abduction. Ethiopia: UNICEF

Hansen, K. M. (2003). *Deliberative democracy and opinion formation*. Denmark: University of Southern Denmark, Political Science and Public Management.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences, 33*(2–3), 61–83.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362.
Janssen, M. A., Holahan, R., Lee, A., & Ostrom, E. (2010). Lab experiments for the study of social-ecological systems. *Science, 328*(5978), 613.
Judd, C. M., & Krosnick, J. A. (1982). Attitude centrality, organization, and measurement. *Journal of personality and social psychology, 42*(3), 436.
Kahan, D. (2000). Gentle nudges vs. hard shoves: Solving the sticky norms problem, 67,University of Chicago law Review 607.
Kamalski, J., Lentz, L., Sanders, T., & Zwaan, R. A. (2008). The forewarning effect of coherence markers in persuasive discourse: Evidence from persuasion and processing. *Discourse Processes, 45*(6), 545–579.
Karpowitz, C. F., Raphael, C., & Hammond, A. S. (2009). Deliberative democracy and inequality: Two cheers for enclave deliberation among the disempowered. *Politics & Society, 37*(4), 576.
Klassen, A. D., Williams, C. J., & Levitt, E. E. (1989). *Sex and morality in the U.S*. Middletown, CT: Wesleyan University Press.
Klofas, J., & Toch, H. (1982). The guard subculture myth. *Journal of Research in Crime and Delinquency, 19*, 238–254.
Kurzban, R. (2001). The social psychophysics of cooperation: Nonverbal communication in a public goods game. *Journal of Nonverbal Behavior, 25*(4), 241–259.
LeJeune L., & Mackie, G. (2009). Social dynamics and the abandonment of harmful practices: A new look at the theory. UNICEF Innocenti Research Centre.
Luskin, R. C., Fishkin, J. S., & Jowell, R. (2002). Considered opinions: Deliberative polling in Britain. *British Journal of Political Science, 32*(03), 455–487.
Mansbridge, J. (1999). Everyday talk in the deliberative system. In S. Macedo (Ed.), *Deliberative politics: Essays on democracy and disagreement* (pp. 211–42). New York: Oxford University Press.
Matza, D. (1964). *Delinquency and drift*. New York: John Wiley and Sons.
Mercier, H. (2011). What good is moral reasoning? *Mind & Society, 10*(2), 131–148.
Mercier, H. & Landemore, H. (2012). Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology, 33*(2), 243–258.
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74.
Miller, D. T., & McFarland, C. (1991). When social comparison goes awry: The case of pluralistic ignorance. In J. Suls & T. A. Wills (Eds.), *Social comparison: Contemporary theory and research*. Hillsdale, NJ: Erlbaum.
O'Gorman, H. J. (1975). Pluralistic ignorance and white estimates of white support for racial segregation. *Public Opinion Quarterly, 39*, 313–330.
Orbell, J. M., Van de Kragt, A. J. C., & Dawes, R. M. (1988). Explaining discussion-induced cooperation. *Journal of Personality and Social Psychology, 54*(5), 811–819.
Ostrom, E. (1991). *Governing the commons: The evolution of institutions for collective action*. Cambridge: Cambridge Univ Press.
Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *The American Political Science Review, 404–417*.
Packard, J. S., & Willower, D. J. (1972). Pluralistic ignorance and pupil control ideology. *Journal of Education Administration, 10*, 78–87.
Platteau, J.-P. (2000). *Institutions, social norms and economic development*. London: Harwood Academic Publisher.
Robinson, C. E. (1932). *Straw votes*. New York: Columbia University Press.
Sally, D. (1995). Conversation and cooperation in social dilemmas. *Rationality and Society, 7*(1), 58.
Schank, R. L. (1932). A study of community and its group institutions conceived of as behavior of individuals. *Psychological Monographs, 43*(2), 1–133.
Stuntz, W. (2000). Self-defeating crimes. *Virginia Law Review, 86*, 1871–1882.

Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy, 10*(2), 175–195.

Tetlock, P. E., Peterson, R. S., McGuire, C., Chang, S., & Feld, P. (1992). Assessing political group dynamics: A test of the groupthink model. *Journal of Personality and Social Psychology, 63*(3), 403.

Thagard, P. (2002). *Coherence in thought and action*. Cambridge: The MIT Press.

Chapter 4
Social Norms from the Perspective of Embodied Cognition

Chris Goldspink

Cognition is not a phenomenon that can be successfully studied while marginalizing the roles of body, world and action.

(Andy Clark, 1999)

Introduction

Like others within this title, I attempt to come to terms with the way in which human social norms emerge from, but are irreducible to, processes at the level of the individual. The particular contribution of this chapter is to suggest an emergentist account of norms which draws on the developing theory of enactive cognition. I use this to consider the characteristics needed for a system capable of simulating human-like norms in a computational environment.

Our understanding of emergence has been greatly expanded through computer simulation, but to date this has cast a light primarily on emergence within physical systems. Attempts to apply lessons from this work to social systems have largely proceeded by attempting to make simple agents more 'intelligent'. The model of intelligence used is generally that of first-generation artificial intelligence—known as cognitivism or representationalism—where the agent is equipped with some

C. Goldspink (✉)
Victoria Institute for Education, Diversity and Lifelong learning, University of Victoria, Melbourne, Australia

Graduate School of Education, Faculty of Social Science and Law, University of Bristol, Bristol, UK

Systems Centre for Learning and Leadership, University of Bristol, Bristol, UK

School of Humanities and Social Science, University of Newcastle, Newcastle, Australia
e-mail: cgoldspink@inceptlabs.com.au

M. Xenitidou and B. Edmonds (eds.), *The Complexity of Social Norms*, Computational Social Sciences, DOI 10.1007/978-3-319-05308-0_4,

limited ability to represent particular characteristics of its environment in a rudimentary computational 'mind' (Franklin, 1998). This has led to many interesting simulations, but these fall well short of allowing us to simulate many of the more complex aspects of human social interaction, including that of norm formation, maintenance and change (Sawyer, 2001, 2005).

Picking up on the statement from Andy Clark cited above (1999),[1] the central argument of the chapter is that much of what is interesting about the mechanisms of norms does not happen between passive agents nor agents with abstracted 'minds' but rather in bodies with brains operating within highly contingent environments, which they themselves contribute to generating and can also potentially change. It is towards understanding this type of phenomena that theories of embodied and enactive cognition are directed. Furthermore, recent advances in robotics have shown how these theories can be applied to practical experiments which in turn support the ongoing development of an emergentist understanding of social behaviour.

I begin with a recount of the central problems enactivist theories are directed at solving. Most fundamental of these is inadequacy of the theorization of the interplay between micro and macro social phenomena. I provide a brief restatement of the contentious issues within emergentism, connecting these to the problem of understanding norms. This discussion includes reprising levels based on the account of emergence which considers the defining features of human social and cognitive agents relevant to understanding normative behaviour. To ground the theoretical discussion I then sketch a typical normative scenario and use it to identify critical cognitive capabilities which appear to play a role in norms. I then provide an overview of the developing enactive account of these capabilities and their relevance to understanding norm emergence. I conclude by comparing alternative simulation paradigms, thereby summarizing where we are with respect to being able to simulate the mechanisms identified as relevant to norm emergence.

The Micro–Macro Problem and Its Implications for Understanding Norms

An adequate theoretical account of norms should pose plausible answers to questions such as the following:

- Where do norms come from?
- How are they maintained?
- What leads them to change or to disappear?

Existing theoretical accounts often fall well short of this in that they lack a sufficiently detailed account of the mechanisms at work. This is surprising as Sripada

[1] Clark is an advocate of embodied cognition; in this chapter I argue more for a more radical extension of the embodied standpoint—that of enaction. The enactive view has it that an agent's cognitive capabilities in effect give rise to distinct worlds—as Varela once expressed it by 'laying down a path by walking'.

and Stich (2006) state: '*No concept is invoked more often by social scientists in the explanations of human behaviour than* "*norm*"'. Indeed, the concept has been incorporated into a wide range of alternative and often competing theories of social behaviour. This lack of agreement about what norms are, and how they operate, has led to the suggestion that it is a generic concept (Gibbs, 1965) with no explanatory value.

The normative literature can be divided into at least two fundamentally distinct perspectives: the social philosophical tradition and the view from the philosophy of law.

In the social philosophical tradition (Lewis, 1969) norms are seen as a particular class of emergent pattern which spontaneously emerge in a population. From this perspective, a 'norm' is identified when a pattern of social behaviour is observed which is apparently prescriptive/proscriptive—people behave 'as if' they were following a rule. This is a bottom-up or a micro-to-macro account.

By contrast, the view offered by the philosophy of law posits norms as a source of social order. This standpoint assumes the prior existence of (powerful) social institutions which are the source of rules. When generally followed, these rules lead to the social pattern we call norms. This is a top-down or a macro-to-micro account.

Clearly the answers to the questions of where norms come from, how they are maintained and what leads them to change mounted from the perspective of these two alternatives differ markedly. However it is possible that there may be a point of synthesis which could unite these apparently opposed viewpoints. This is what an emergentist account of norms attempts, and it is linked to wider concerns about the relationship between micro and macro phenomena within the social sciences. It is worth revisiting where the debate about the micro–macro problem currently stands before attempting an extension of that debate with specific relevance to understanding norms.

Simply stated the problem is that we have no adequate way of accounting for the relationship between the (bottom up) actions of individuals and resulting social structures and the (top down) constraint those structures place on individual agency. This is not for want of trying.

The problem is central to many nineteenth- and twentieth-century social theories. Examples include Marxian dialectical materialism (Engels, 1934) built upon by, among others, Vygotsky (1962) and Leont'ev (1978); the social constructionism of Berger and Luckman (1972); Giddens' structuration theory (1984) and the recent work of critical realists (Archer, 1998; Archer, Bhaskar, Ciollier, Lawson, & Norrie, 1998; Bhaskar, 1997, 1998). These alternative theories are frequently founded on differing assumptions, extending from the essentially objectivist/rationalist approach of Coleman (1994), through the critical theories of Habermas and the radical constructivism of Luhmann (1990, 1995).

Many of these accounts conclude that structure and agency come together in *activity* or in *body-hood*—the specific psycho-motor state at the instant of enaction. Both Vygotsky and Giddens, for example, focus on action as the point of intersection between human agency and social structures, and it is implicit in Bourdieu's *habitus* also.

Essentially all of these accounts have the limitation that they fail to provide an account of the mechanisms which link the micro and macro conditions in a way which can be tested empirically or be made operational such as through multi-agent simulation. I will return to this challenge in the final section of this chapter.

In the recent past, our understanding of the mechanisms that connect micro and macro has been significantly advanced by the systems sciences, and in particular complex systems theory, as well as by developments in social simulation, evolutionary robotics and artificial life. Much of this has been done under the banner of emergentism, and this is the viewpoint that I will be developing here.

What I am essentially arguing in this chapter is that to provide an adequate account of norms we need an adequate account of *social* emergence. The key challenge is that mechanisms of emergence are likely different when we consider natural systems and social systems. We recently provided a brief account of the history of the concept of emergence and its contribution to current thinking about the interplay between micro and macro social phenomena and suggested a form of emergence particular to social phenomena which we called reflexive emergence (Goldspink & Kay, 2007; Goldspink & Kay, 2008). Reflexive emergence is associated with agents, such as humans, whose cognitive capabilities make them self-aware and strategic in their actions. This work is built on earlier contributions to understanding different orders of emergence—the argument that different cognitive capabilities support qualitatively distinct forms of emergence.

Orders of Cognition Give Rise to Orders of Emergence

Gilbert (2002), for example, has distinguished between what he called first- and second-order emergence. First-order emergence includes macro structures which arise from local interactions between agents such as particles, fluids and reflex action. This corresponds to the focus of interest to natural scientists and much of the research into complex systems which has its origins in the natural sciences. Second-order emergence is argued to arise '*where agents recognise emergent phenomena*, *such as societies*, *clubs*, *formal organizations*, *institutions*, *localities and so on where the fact that you are a member or a non-member*, *changes the rules of interaction between you and other agents*' (Gilbert, 2002: 6).

In a similar vein, Castelfranchi has distinguished what he refers to as cognitive emergence which '… occurs where agents become aware, through a given "conceptualization" of a certain "objective" pre-cognitive (unknown and non deliberated) phenomenon that is influencing their results and outcomes, and then, indirectly, their actions' (1998: 27). Castelfranchi thus conceives of a feedback path from macro pattern to micro behaviour and specifies a cognitive mechanism. He argues that this mechanism has a significant effect on emergence and gives rise to a distinct class of emergent phenomena. These ideas are more comprehensively reflected in the five orders of emergence suggested by Ellis (2006: 99–101). All these argue that the range and type of emergence possible in a system depend fundamentally on the

range and class of things agents are able to distinguish and the behaviour they are able to generate. If we are concerned to understand the emergence of norms in human societies we therefore need an adequate account of human cognition and the different aspects and facets and how they play a role in norms.

There has been considerable research directed at understanding the origins and developmental phases associated with distinctively human cognitive capabilities. Much of this has drawn on comparative neurology and sociological and psychological study of non-human animals, in particular apes. Insights are available also from the developmental psychology and neurology of the phases of development from infant to adult (see for example Reddy, 2008; Smith, 2005; Smith & Thelen, 2003).

Gardenfors (2006) identifies the following as among those needing explanation (presented in order of their apparent evolution): emotions, memory, thought and imagination, self-consciousness/theory of mind, free will and language. These are present to varying degrees in different organisms and develop at different stages in humans from infancy to adulthood. The degree of interrelatedness is not, however, straightforward. Apes for example demonstrate self-awareness and 'theory of mind' but do both without language, whereas in humans language appears to play a significant role in both.

Which of these cognitive capabilities are implicated in norms and in what way are considered briefly below and then developed throughout the rest of the chapter.

Cognitive Capabilities Implicated in Norms

Therborn has argued (2002: 868) that people follow norms for different reasons. He argues that at the more limited end of the range this involves habit or routine. Considering the cognitive capabilities implied in this, a simple capacity for remembering would be sufficient. He also argues that rational knowledge of consequences for the world may be involved. This implies agents capable of consciousness and free will or agency. Considering the implications of the previous discussion of orders of emergence, there may not, therefore, be a single emergentist account of norms, but rather a family of related ones. In other words, the overgenerality of the concept of norms may have led to a range of social behaviours being grouped together where very different generative mechanisms are implied. This is an important point from the perspective of this book as while those aspects of norms which are associated with memorised actions or unconscious patterning in decision making may lend themselves to being modelled with current approaches to social simulation, those which depend on conscious awareness do not. I say 'may' as recent research shows just how intertwined the evolutionarily older and more recent cognitive capabilities are in humans, where reflex, affect and rationality play out in complex ways in decision making (Lehrer, 2009); our physical experience of being 'in the world' informs our cognising and reason (Johnson, 1990; Lakoff & Johnson, 1999) and emotion permeates and is central to 'rational' action (Damasio, 2000, 2006).

In summary, understanding the mechanisms underpinning norms implies coming to terms with some of the most vexing aspects of social science: the problem of structure and agency or the micro–macro problem. Advances in our understanding of emergence have been driven by recent developments in complex systems as well as by the many and various examples of simulating both natural and social phenomenon. Through this work we have increasingly come to understand that different agent capabilities potentially give rise to different orders of emergence and that examples of emergence in the natural or the animal world may not help much with understanding how emergence works in human social systems—including norms.

A Narrative Account of Norms to Ground the Theoretical Discussion of the Role of Cognition

To explore the cognitive capabilities potentially involved with norms further, as well as to provide real-world grounding for the necessarily abstract discussion which follows, let us take a simple hypothetical narrative account of the operation of norms. In this simple narrative I will incorporate interactions which may play a role in the three questions with which I opened this section:

- Where do norms come from?
- How are they maintained?
- What leads them to change or to disappear?

I am a foreigner recently arrived in a new country. Walking the streets I follow the norm of my culture which is to acknowledge the presence of those (including strangers) I encounter. This pattern of engaging is for me habitual and unconscious. Let us assume that on first doing this my attempt to engage is ignored and gaze averted. My protagonist may also be acting out of habit. How do I know if this habit is based on a social norm and one that has salience to me? At this stage I do not and I may never do so, and yet I may still participate in maintaining or disrupting it.

On this first encounter, if I think about the reaction of my protagonist at all, I may conclude that he/she is simply acting out of an individual disposition—shyness perhaps. From his/her perspective I may be perceived as brash or threatening—again acting out an individual disposition rather than following a norm. The encounter may have registered unconsciously—we humans, like many animals, have evolved acuity to detecting patterns in behaviour which are contrary to our expectations (Lehrer, 2009). When I acknowledge my protagonist I may have triggered a physiological reaction. This may have included a tensing of the body, pulling away from me and the aversion of eyes as well as micro gestures of the face which suggest aversion—perhaps a flicker of shock or fear (Ekman, 1992). I may perceive these unconsciously at first through the somato-visceral system—I may experience negative affect, and I may become conscious of it as a feeling of surprise. This too may be unconsciously signalled through micro gestures although my protagonist may

not notice as he/she has already averted gaze. For both of us the reaction has 'meaning', and in the most general level this is one of threat.

If I had become conscious of the encounter I may describe it as having been rebuffed. My protagonist may report having been threatened. However, at any stage, neither the encounter nor the reaction enters conscious awareness.

Let us assume that over successive days the experience is repeated with different individuals. The negative affect experienced may lead me to unconsciously adjust my behaviour—I become less forthright or even mimic the response in order to re-establish a pattern that avoids the negative affect. If I mimic, and the response I encountered was indeed based on an individual disposition and did not reflect a social norm (there was a half-way home for paranoids nearby), then I may begin a norm as I now avert my eyes from even the non-paranoid and potentially change their behaviour. If avoidance was already a norm, then I now successfully contribute to its maintenance.

If at some point the interaction does enter my consciousness then a wider range of responses becomes possible. I may decide that the nationals of this country are antisocial and decide to 'play' with them, for example. I become even more intrusive—verbally greeting people to delight in their discomfort (rather like turning and facing people in a lift). Alternatively I might come to appreciate that this is a norm, but one particular to this place or to certain people within this place. This may help me be more tactical in the way I behave, choosing alternative ways to interact based on my appraisal of the situation and what I want to gain from my interactions with the others present. Over time this may become unconscious again—I hear a certain accent and I avert my gaze, a native of my own country, and I fully gesture acknowledgement.

In this account the degree of entanglement of cognitive abilities is illustrated. A norm may be effectively initiated or maintained without conscious awareness with signalling of conformance or non-compliance happening through subtle micro gestures out of awareness of one or more of the participating individuals but, equally, may be influenced by fully conscious processes. It may or may not involve deliberate action and consideration of own or others goals, interests or needs.

The encounter may only ever involve dyadic exchange—me and a particular protagonist. In that context neither of us can say anything about the presence or the absence of norms as we both lack the wider perspective to judge the behaviours as shared. The encounter cannot be understood without an appreciation that both of our reactions are the consequence of many past interactions which each of us has had within two different social contexts leading to the establishment of habits of action which maintain our social acceptance within that particular social context. Nevertheless, the social context determines what happens next.

If I am in a social context for which the habit is non-adaptive then the succession of disconfirming interactions and the affective impact this has on me will likely lead me to adjust my behaviour. Over time a new accommodation may be reached. If we were to go and seek out first-hand accounts of the experience of the encounter we would find very different attributions. I may describe being 'rebuffed by an antisocial person'. My protagonist may describe having been 'accosted by a foreigner'.

These accounts need not play any role in the process but they may. In making such an attribution I may decide to undertake a campaign to deliberately act so as to 'socialise the locals'. I use my agency to amplify my behaviour when I judge that it may be effective. In so doing I may generate even stronger reactions and deepen the norm among those I seek to influence—an unintended consequence.

Alternatively the rejection may lead me to give up and go and find more people whose self-narrative I share (hang out in expatriate pubs). Alternatively, within my social circles at least, I may succeed and over time this may propagate beyond my immediate interactions and change the established norm. All of which is to say that no individual needs to be aware of the 'norm' as norm, nor agree or consciously follow the deontic implied, indeed may even consciously and deliberately refuse to follow it and yet will participate in the maintenance (or potential change) of that norm. Whether or not the norm is maintained or changed will depend on the current state of the social system as a whole—including such factors as relative number of 'followers' compared to 'challengers' and how they have self-organised (distributed compared to ghettoised), the rate of introduction of individuals not accommodated to the patterns of the dominant social group. All of which is to argue, in the loosest possible way, that norms are indeed emergent. The challenge then is to more rigorously theorise what we can readily recognise.

Theorising the Mechanisms of Human Social Norms

Based on what has been presented above, the key point I wish to develop in this section is that both the prior social emergentist theory as well as a simple narrative account of norms in action imply that human social norms involve agent cognitive capabilities of various types operating at multiple levels. We need a theoretical account which can synthesise this into a framework which is compatible with an emergentist perspective and which can support practical experimentation and empirical investigation. I argue here that an enactive view is the best theory we currently have for this even though it is very much a work in progress and brings its own challenges.

In the remainder of this chapter I first provide an overview of key developments in an enactive theory of cognition and then examine the implications this has for the empirical study of norms as well as for their simulation.

Towards an enactivist account of norms

In the narrative encounter described above it is apparent that the history of past interactions in a particular social domain influences how each individual behaves instant to instant. This is consistent with the theoretical idea distilled from the many past attempts to come to terms with the interaction between micro and macro levels: structure and agency come together at the point of enaction. The fact that it is automatically reflected in all aspects of the agent (somato-visceral, affective and sensori-motor) indicates also that we are not talking just about deliberate action but

states of bodies as well as brains. For Bourdieu the habitus was the embodiment in each individual of the past as '*dispositions*, *schemas*, *forms of know-how and competence*'. For him also these were effective due to their being '*below the level of consciousness and language*, *beyond the reach of introspective scrutiny or control by the will*'. In discussing Bourdieu's account Crossley (2001: 83) states that, as a consequence, what was sought is '… *a conception of human action or practice that can account for its regularity*, *coherence*, *and order without ignoring its negotiated and strategic nature*'.

The construction of such an account has begun. It is being informed by developments in evolutionary biology, cognitive science, neurophysiology, robotics, artificial intelligence, artificial life as well as psychology, social theory and philosophy (Stewart, Gapenne, & Di Paolo, 2010). It represents an ambitious program to unite currently disparate perspectives on what it is to be an autonomous and intelligent agent. The wide scope of this enterprise presents a challenge in the context of this chapter: how best to summarise current development and link it to the theme of norms. Recent work by Barandiaran and Di Paolo et al. (Barandiaran, 2005; Barandiaran, Di Paolo & Rohde, 2009) as well as by Damasio reinforces a key theme—that the higher order abilities implicated in norms rest on the fundamentals of our living being and so we have to begin with biology, albeit emergentist biology.

The biological origin of what is meaningful and what is 'good' and 'bad' for an agent and therefore of what it 'ought' to do

In the account of norms provided by the philosophy of law discussed briefly in the opening section of this chapter, it was noted that norms imply a deontic—what 'ought' to be done. From this theoretical account the deontic is supplied by the wider society or by powerful social institutions within that society. This is in contrast with a dialectical account in that it provides no explanation of how such institutions come to take on significance or authority—to be meaningful—from the point of view of the individual, nor why individuals accede to them. The enactive account shows how this can come to be, and yet how the deontic has its origins in biological fundamentals. The account is a radical departure from how we habitually think about such things, and the following account may appear quite circuitous. It is necessary, however, to explain how some phenomenon comes to have 'meaning' for the agent.

The Biological Basis for Meaning

The transition between living and non-living has been argued, in emergentist terms, to result from self-organisation—more particularly a chain of autocatalysis resulting in the formation of self-producing autonomous (autopoietic) entities (Maturana & Varela, 1980). Recent extensions of this theory (see Barandiaran, 2005) have it that a minimal cognitive agent has a primary metabolic loop which serves to maintain its

biological viability and (at least) one other loop which links sensory surfaces with motor surfaces. This second loop adds significant plasticity within a behavioural rather than a metabolic domain (Moreno & Etxeberria, 1995: 168). Approached in this way '*minimal cognition is not so much a centralized property of the biological hardware of an organism, …*' as many theories of mind would have it, '*or a set of internally computed algorithms, …*' as assumed by first-generation artificial intelligence, including that which underpins much contemporary social simulation, '*but instead denotes an abstraction of organism environment reciprocity*' (van Duijn, Keijzer, & Franken, 2006: 165).

The most important implication of this is that the agent's classification of, and accommodation to, its environment is dynamic/homeostatic. Rocha uses the language of complex systems to elaborate on this, arguing that the order or the stability implied in the maintenance of agent viability—autopoiesis itself—is an attractor, as are the various metabolic and sensori-motor cycles involved in maintaining its relationship to a dynamic environment. States on these attractors constitute sources of input or reference to other attractors, and the current configuration of the nested attractors tells us something about the agent/environment accommodation at a particular point in time. As Rocha states it, these '*… perform environmental classifications … not all possible distinctions in some environment can be grasped by the self-organizing system: it can only classify those aspects of its environment/sensory motor/cognitive interaction which result in the maintenance of some internally stable state or attractor*' (Rocha, 1996). In other words, the range and type of environmental triggers that can be accommodated by an agent are necessarily constrained by the agent's biology, physiology *and* ontogeny and are reflected in its dynamical structure at any given point in time.

Importantly, those triggers which lead to a compensatory action can be said to be 'meaningful' from the perspective of the organism in that they have implications for its state and viability—what is 'good' for it or 'bad' for it—and may link directly to reflexes which serve to orientate it towards the 'good' (follow a nutrient gradient towards a source) and away from the 'bad' (move from an area of excessive or insufficient temperature). This is consistent with the position taken by Varela (Rudrauf, Lutz, Cosmelli, Lachaux, & Le Van Quyen, 2003; Thompson, 2004; Varela 1997) that what agents are sensitive to is determined by their own operation, not the environment. This establishes conditions of relative autonomy in that '*It is not the organism that matches the environment in a given specified way. On the contrary it is through the particular way in which the agent satisfies the homeostatic maintenance of essential variables that an adaptive environment* (*a world*) *is specified—cut out from a background of unspecific physical surroundings*' (Barandiaran, 2005).

However, this description of simple autonomy is still a long way from issues of higher cognition and norms. It is this connection I discuss next.

The idea that agent states define what is meaningful to them has direct parallel to the concept of affordances in social theory (Gibson, 1977). Particular organisms are capable of distinguishing particular stable structures in the environment, and these structures, when combined with the organism-specific capabilities, 'afford' those

organisms some opportunity. Looked at another way, material features of the world become tools, artefacts and technologies for that organism to the extent that they can extend that agent's cognitive range. In social systems also, existing social structures 'afford' opportunity and facilitate certain actions, extending the cognitive range and action potential of individuals. In part the argument here is that an agent's cognitive boundary may not be co-extensive with its physiological boundary.

With the account provided so far, we are building a layered model. The sensori-motor loop/s associated with a class of agent supports (support) distinct 'phenomenal domains'. These domains are loosely coupled to each other and to that generated by the metabolic processes associated with autopoiesis. In other words each domain has its '… *own internal coherency*' which constitutes a '*meaningful world in itself*' (Barandiaran & Moreno, 2006: 176). A social illustration of this partial autonomy or loose coupling of domains is the recent phenomena of suicide bombers. Here 'meaning' in one phenomenal domain (the belief in paradise) can trigger a behaviour which is inconsistent with the fundamental operation of the metabolic phenomena needed to maintain life. The organism is destroyed as a result of the operation of mechanisms which otherwise serve to extend and maintain its viability through inclusion within a particular social domain—in this case mutual acceptance around the norm of belief in fundamental precepts of a religion.

As we add layers of sensori-motor loops we need something to integrate then—a central nervous system. The advent in evolutionary terms of central nervous systems does not change the account of cognition provided so far in any significant way. Cognition does not now happen in brains: it is still in the agent/environment interaction. What is meaningful is not stored as a representation in memory; it is still in the dynamic maintenance of viability operating between the agent and its environment. All that has changed is that now this is facilitated by the nerve systems which link expanded points of interface with that environment. We can now say that it is the nervous system's structure—by which we mean the attractor states established within it rather than its physical architecture—that dictates which environmental perturbations can be a trigger (Mingers, 1991; Varela, Thompson, & Rosch, 1992) and therefore what will stand in a 'meaningful' relation to the agent. Just as with the amoeba, this has the implication that, as each organism traces a unique history, it specifies what is meaningful to it within its environment. Agents which trace similar or even share histories will generate similar domains of meaning (similar things in the environment as well as in the behaviour of each to the other will carry similar implications for their respective viability), while those which trace very different histories with little or no sharing may generate unique domains of meaning. We saw an example of this in a hypothetical human system with the two sets of cultural norms present in the narrative of a foreigner in a new country.

We may already talk about patterns in these resulting accommodations as 'norms' even if they are only coordinated by simple and largely innate reflex actions. Norms then are shared domains of 'meaningful' accommodations between agents. What is meaningful, and indeed the meaning conveyed, is referenced ultimately to that which is essential to maintaining the viability of the agent.

The Role of Affect and Emotion

A significant change in cognitive theory over recent times has been the growing acceptance that emotion is fundamental to cognition, including that of humans (Colombetti & Thompson, 2008; Damasio, 2006). This work suggests a complex relationship between aspects of the functioning of the body and is consistent with the intertwining of cognitive capabilities discussed earlier.

From the perspective of contemporary research, affect is argued to provide a rapid primary appraisal of presenting situations which operates in advance of conscious categorisation or assessment: affect directs the attention of the agent towards aspects of the environment or its own state which are relevant to its viability. This 'core affect' is argued on the basis of considerable empirical evidence to be a relatively un-differentiated state of arousal measured by the dimensions of valence (good/bad) and arousal (activated/deactivated) (Ryan & Deci, 2001). '*Core affect has been characterized as the constant stream of transient alterations in the organism's neurophysiological state that represents its immediate relation to the flow of changing events*'—*it is* '*a neurophysiological barometer of the individuals relation to an environment at a given point of time*' (Barrett, 2006: 31). Affective states then afford to an animal capable of supporting them what a simple sensori-motor reflex did for the amoeba, a means for classifying states in the agent/environment interaction as 'good' or 'bad'. Negative affect becomes associated neurologically with past experiences and conditions which were harmful and positive affect with ones that were beneficial to the agent.

What we call emotion is built on this core affect. Barrett (2006: 25) argues that '*The taxonomic structure of self-reported experiences of emotion does not support the view that anger*, *sadness*, *fear and so on*, *are qualitatively distinct and experientially primitive*'. This is to say that emotions are not biologically primitive like core affect but arise from a process of conceptual or perceptual categorization on top of or in relation to an affective response. What we commonly refer to as emotion (or *feelings* in Damasio's schema) are labels for a set of experiences represented in consciousness.

The position taken by many of these more recent emotion theorists is that these 'conceptualisations' are not abstracted from sensori-motor events and stored in propositional form, but exist as 'simulations' ('as if' states for Damasio) of the sensori-motor states that occurred with previous instances of a similar experience. When we see a picture of something frightening, we do not recover an abstract concept of fear to label the picture, rather we re-experience fear at a somato-visceral level, albeit in a low key way—the concept of fear is embodied.

Affect then represents a whole body state response to environmental triggers. Even when triggered by memories of events, they elicit a response that involves arousal and action—affecting the viscera, endocrine and motor systems in concert. This then presents no problems from the account of cognition being presented—affect and emotion merely form part of the continuum which may support qualitatively distinct domains of interaction and hence mechanisms for norm formation

and maintenance. A possible role for emotion in norm forming, maintenance and change was indicated in the narrative of a foreigner in a new country provided above. In this account, affect played a significant role in the immediate flow of events. Subsequent reflection on the emotional experience could have effected subsequent interactions to the extent that embarrassment was experienced in suffering a rebuttal to a social exchange, in endeavouring to avoid the experience in future. Each time the event is remembered, by the above account of the operation of emotion, the associated affective state will be re-experienced, serving to deepen the experience and aversion and perhaps the resolve to behave in some different way. For me to resolve to do something, however, I need to have conscious self-awareness and perhaps a sense of identity about who 'I' am as well as agency. These aspects too then can (but may not) play a role in the creation, maintenance or destruction of change of norms.

The 'Viability Set' Provides a Foundation for What Is 'Good' or 'Bad' at the Level of the Individual

Building on what has been argued so far we can say that 'cognitive agents' define what is meaningful to them in the environment—they place value on the stream of events they experience as they experience them. For living agents, at the most fundamental level, what stands as meaningful are those aspects of the environment essential to maintaining their viability as a living entity. For the most basic organisms (such as cells) their ability to adapt and remain viable in response to a change in their environment is quite narrow—specified by chemical and mechanical parameters fundamental to their metabolic pathways. However, once an organism has developed a sensori-motor loop in addition to the purely self-maintaining metabolic mechanisms it has the capacity to adapt behaviourally to its environment. Simple amoeba can, for example, propel themselves along a nutrient gradient using simple mechanisms such as flagella. This response capability is bounded: sensitive to only a limited range of changes with a limited set of response capabilities (flagella only work in fluids of limited range of viscosity). We can therefore conceive of a 'viability set': the range of events to which the organism can adapt and maintain its viability (Di Paolo, 2005). The basic sensori-motor mechanisms of reflex through to affective pre-appraisal (as just discussed) and then reaction through to conscious decision making and language (which I will consider in the next sections) all serve to expand the viability set.

As an agent begins to interact with others the response it engenders will be perceived as affirming or as a threat. With human agents this will most likely initially take place based on affective pre-appraisal (Damasio, 2000) as discussed above but may also involve more conscious deliberation as included in the account of the foreigner in a new country. The evaluation will lead to a behavioural response which is adaptive—based on the agent's history of interaction in particular social

domains in its history. The implicit goal will be to increase the chance of remaining viable in the current domain. It does not, however, do this in splendid isolation.

The discussion so far has focused on individual cognitive capacity—a very micro focused orientation. However the moment the effect of coupling between micro-agents is appreciated the pathway and mechanism by which social structures bootstrap from these interactions and back-propagate to constrain them become apparent. This next step is therefore key to an emergentist account of norms.

As agents interact with one another, their viability sets intersect. We could model this in the same way Kauffman (1993, 2000) has for fitness landscapes. The resulting 'viability landscapes' are coupled—the adaptations made by one agent change the landscape of the others with which it is interacting. In Froese and Di Paolo's terms '*… since the regulation of the interaction of one agent changes not only its own coupling but also that of the other agent, it follows that the agents can enable and constrain each others sense-making*' (2009: 9).

At the most general level norms can be conceptualised as relatively stable patterns on this coupled viability landscape—agents converge on viable accommodations of each other's accommodations. They form from the complex product of the response capability of the agents—affective, unconscious as well as rational conscious, but where each agent influences others through its behaviour (which may include subtle gestural aspects as well as the more overt). In this sense, norms are possible as agents make mutual accommodations to one another so as to maintain their viability within a particular social domain. However, if we are to make sense of behaviour such as the ignoring of norms then the mechanism described so far, that of viability maintenance, is insufficient. We need another idea: that of agency.

Agency

In considering a the role of agency in norm formation and maintenance we are concerned to distinguish between purely adaptive accommodations to environmental change, including that generated by the action of other agents in the coupled viability landscape, and agents which modulate their own behaviour so as to shape the trajectory of their interaction with the environment. Barandiaran et al. discuss it as follows:

> Environmental conditions are good or bad for the continuation of the system. This normative dimension is not arbitrarily imposed on the system by a designer or external agent that monitors the functioning of the system and judges according to her interests. It is the very organisation of the system which defines a set of constraints and boundary conditions under which it can survive. … This precariousness implies that whatever the organism is doing … there is something that it ought to do; not for an external observer but for itself, for the continuation of its very existence (2009: 375).

This quote illustrates why we had to go back to discuss fundaments of biology in order to understand norms. What is 'meaningful' to an organism and hence the base

for all subsequent accommodations and judgements about what is 'good' or 'bad' for it propagate from its biological viability. It will now be apparent to some readers that this presents a problem from the point of view of simulation. I will have more to say on that in the final section.

Barandiaran (2005) has argued that loose coupling between the metabolic domain and the sensori-motor domain allows an organism to exploit the rapid response times of the neural system in order to expand its viability set. Within the emerging field of neurodynamics (Cosmelli, Lachaux, & Thompson, 2007; Kelso, 1995; Rocha, 1996; Thompson & Varela, 2001; van Gelder, 1998) it is argued that this 'plasticity' is in large part due to the nervous system operating on a system of complex attractors, yielding quasi-stable emergent states. By these accounts it is the asymmetry between the combination of all possible configurations the agents biology and ontogeny afford it, and the (more limited) range of responses needed to maintain immediate regulation in a given environment, that gives rise to what we call 'agency': '*The higher the agent's capacity for adaptively guided self-restructuring* (*plasticity*) *the higher its behavioural adaptive autonomy and hence its agency*' (Barandiaran, 2005).

Peter Hejl (1993) also locates agency in 'cerebral overcapacity'. He notes that this conveys advantages and disadvantages. The advantage is in furnishing support for a wide range of possible responses and hence 'requisite variety' (Ashby, 1974). The disadvantage is that high plasticity contributes to the contingent nature of agent–agent and agent–environment interactions and thus instability. The advantages only hold sway over the disadvantages to the extent that the variability can be channelled or constrained in short time frames. As Hejl notes, '*The only* '*solution*' *to this problem … seems to be society*' (1993: 229). For Hejl then quasi-stable structures that arise through social interaction (such as norms) serve to reduce social complexity in the short term while keeping open a much wider range of possible adaptations and accommodations—through the change of existing norms or emergence of new ones appropriate to alternative contexts.

In short then, the 'surplus capacity' made available by an advanced neural system explains how a living system can come to have the potential to remain viable in changeable environments, but not how it exploits that potential. There is still a perspective missing. This is the perspective of how an agent can come to be conscious of its capacity for choice and use that choice in strategic and tactical ways.

As humans we can choose to ignore a norm—perhaps rationalising that it does not apply to us. As was illustrated in the narrative of a foreigner in a new country, we can also choose to maintain or to try and change a norm or begin a new one. All of this implies the use of agency in a strategic way—a purposeful striving. This only becomes possible if the agent can distinguish 'self' from 'other' and can act to advance its own or others' interests and intentions in a deliberate, selective and conscious way. Consciousness needs to be explained as a higher order cognitive function with significant potential implications for normative mechanisms, and, in the context of this chapter, it also needs to be placed within the wider enactive account being developed.

An Enactive Account of Consciousness, Self-Awareness and Identity

Thompson has argued that the sense of 'self' has its primary (pre-conscious) origin in an organism's capacity to use its own self-constitutive processes as a source of reference. Here the sense of self as a 'totality' or a stable whole is strongly associated with its biological autonomy (Thompson, 2005) and hence has its origin in fundamental biological processes such as those already discussed above. Similarly, Damasio (2000) distinguishes between proto-self, core consciousness and extended consciousness with each being developed on the former. The proto-self relies on the nervous system's capacity to use relatively stable internal states as a reference point. Damasio groups them under the heading of the 'internal milieu'. However, he also argues that this sense may be combined with proprioception and kinaesthetic mappings which identify the positioning of muscles and limbs in combination with the sense of 'fine touch' from the epidermis and thus use the body's interaction with the environment as a reference point for a sense of self as separate from environment. Either may provide a source that is relatively stable which can be used as a foundation for a distinct sense of 'self'. Importantly these sources are always available while the organism is alive and interacting in its environment. This is argued to provide a basis for consciousness to the extent that the organism can notice that actions have 'self' as an origin (are 'owned' by self) and that through such actions 'self' exerts agency on the environment.

This sense of self is further differentiated. Damasio uses the terms core consciousness and extended consciousness, while others refer to it as minimal self and narrative self. The former is associated with the agent's '*consciousness of oneself as an immediate subject of experience*, *unextended in time*' (Gallagher, 2000: 15) and the latter '*A more or less coherent self* (*or self-image*) *that is constituted with a past and a future in the various stories that we and others tell about ourselves*'. It is only this last form of 'self' or identity construction that requires language. As Menary argues '*First there are the experiences of a living body and then we turn those experiences into a narrative*' (Menary, 2008). Through narrative, however, a variety of alternative stories about self may be elaborated.

Narrative represents a means by which some socially located stability, such as 'norms', capture, propagate and give persistence to the unfolding dynamics of social interaction. They constrain individual action through their shaping of identity, without the individual having permanently to give up the full potential of the wider space of possibilities. They serve to smooth the otherwise turbulent 'push' and 'pull' of the accommodations individuals need to make to remain viable in different social domains. And, in so doing, they may stabilise the wider dynamics that results from structural coupling: forming another layer of constraint on the coupled viability landscapes already discussed.

At the level of the individual, the current state of their ontogeny is reflected in their narrative account of themselves at that time. That narration also reflects their location of themselves in a shared or a social history. Ochs and Capps state, '*The power*

to interface self and society renders narrative a medium of socialization par excellence' (Ochs & Capps, 1996: 31). Returning to the hypothetical story, when our foreigner jokes about the 'locals' with fellow expatriates he or she perhaps construct a narrative which locates 'us' as 'together against another'. The narrative reinforces the shared valuing of one set of norms (the ones shared with those present) and deprecates those of the 'other'. These exchanges, while undertaken in language, invoke emotive responses which become attached to the labels of 'us' and 'them' and will be regenerated in subsequent encounters, influencing behaviour.

While the proto-self is grounded in affect, the narrative self implies language and I have not yet accommodated language into the unfolding account of the relationship between cognitive capability and social norms.

Cultural Tools and Language

Ross (2007: 718) says of language, '*similar public linguistic representations cue similar behavioural responses in individuals with similar learning histories*, *as a result of conventional associations established by those histories*'. Thus, as Maturana has argued, a shared history of interaction leads to the establishment of a consensual domain (Maturana, 1978; Maturana & Varela, 1980). However, contrary to the conventional assumption this does not imply that language constructs a one-to-one denotative representation with objects or phenomena in the real world (Kravchenko, 2007). Rather language represents a particularly flexible form of behaviour by which one agent may attempt to influence another or others. If we concentrate on how people attempt to influence each other in language we will notice that it is not only, or even so much, the content of what is said that matters but more the manner of the saying and hearing. Linguistic interaction cannot be decoupled from the behaviour of talking and listening. Individuals are orientated to one another, and the reciprocal behaviours associated with a stream of 'communication' present each participant with many cues, some more subtle than others, about the others' orientation and intent with respect to the 'self' as well as their apparent purpose and what they intend for and from you. For Cowley and Macdorman (2006) talk is better approached as '*… a multimodal way of toying with persons*'.

If language is more indexical rather than symbolic, utterances and words, as well as the tone and style by which they are delivered, rely on some level of experiential grounding—a learned association gained through repeated exposure within a shared social domain. In this context a word is indexical of a gestalt of sensori-motor experience initially associated with particular contexts but which may become more generalised through increasingly diverse association. Lakoff and Johnson (1999) argue that this is so profound that many of our fundamental concepts 'borrow' from our experience in physical space. So when I say that to perform a task is 'below me' I use a physical metaphor (my experience in the world of things which are above and below one another) to tag an affect which cues me to my place within a social status norm within the society to which I belong. It is this fundamental

characteristic of language which supports the wide range of ways in which we use it—as metaphor, to invoke paradox, to hint at associations, ironically, to provoke, to stimulate and to frustrate, making it a powerful tool for influencing the behaviour of others and hence shaping the formation and transformation of norms.

Language too then plays a fundamental role in modulating shared viability sets.

Part 3: Implications for Simulation of Social Norms

I have now set out the key elements of an enactive account of aspects of human cognitive capability which may play a role in the initiation, maintenance and change of norms. The above account integrates existing psychological, sociological and cognitive theories of human action. It is also consistent with an emergentist approach applied to social systems. The account is far from complete however. It has drawn on recent developments in all of the contributing disciplines, including evolutionary robotics and artificial life and also some aspects of social simulation. It also has the potential to guide these more empirical sciences of sociality. It is to this that I wish to turn in concluding this chapter. In this final section I unpack the implications for how we might approach the simulation of norms mindful of what the enactive view has suggested as key mechanisms.

Our insights into and ability to theorise about the micro–macro interplay at the core of social phenomena have been greatly advanced by the possibility for computer simulation. Much of what we now understand about the behaviour of emergent systems has resulted from simulations. Theory and modelling have therefore moved hand in glove, and we might reasonably expect this to continue. The account set out above has a number of implications for how we choose to model and how we compare the model to the world.

Alternative Paradigms

There are a number of alternative ways in which simulation is being used to advance our insights, particularly into human social system behaviour, including that of norms. The three primary (paradigmatic) approaches are cognitivism, embodied cognition (Clark, 1998; Shapiro, 2011) and, more recently, enactivism (Stewart et al., 2010). Each represents a logical progression in that each is argued to address limitations and problems of those which have come before.

The message from the story recounted earlier is that the regularity which characterises norms is a product of contingent, situational specific striving of the participating agents, acting through a variety of motives, interpreting their situation differently and pursuing a mix of individual and collective goals with each influencing the other on a coupled viability landscape. If this is accepted then we can use this to examine which of the alternative paradigms may support simulation methods best equipped to deepen our understanding of different aspects of normative behaviour.

In this final section, therefore, I want to evaluate where we are in the development of alternative approaches to understanding and modelling norm-capable agents and how we might best advance theory and experimentation directed at better understanding how norms arise, are maintained, change and disappear.

Paradigms of Mind

As previously discussed, simulation including of simple 'dumb' agents or particles has given us a great deal of insight into mechanisms of emergence and will no doubt continue to do so. However, in order to extend this learning into the mechanisms of social behaviour we have needed to make assumptions about the nature of social agents. More particularly we have had to find ways to construct agents which reflect the cognitive capabilities associated with human social behaviour.

The science of artificial intelligence as well as of multi-agent systems has built upon cybernetics which itself drew on information theory and theories of universal computation to posit intelligence as a form of computation. The resulting paradigm has been labelled representationalism or cognitivism. In their book *The Embodied Mind*, Varela et al. (1992), argue that '*The central intuition behind cognitivism is that intelligence—human intelligence included—so resembles computation in its essential characteristics that cognition can actually be defined as computations of symbolic representations*' (Varela et al., 1992: 40).

Cognitivism therefore constructs a duality. The environment is experienced as a 'fact' external to the agent and is acted upon directly but is also conceived and symbolically represented in the 'mind'. This approach gave rise to two well-known and fundamental problems now referred to as the framing problem and the grounding problem. Both of these are relevant to understanding and simulating norm emergence.

As has been discussed, people unconsciously or consciously follow norms on some occasions and not on others. Within social theory this is usually explained by norms being context specific and by agents weighing the cost of adhering to norms against other alternative goals or drives. It is in relation to this aspect of norm following that the framing problem is an issue. Systems based on cognitivism cannot deal with dynamic and subtle variations in context. They require the designer to anticipate the range of environmental conditions the agent will encounter and design in a set of decision rules to support this.

The grounding problem is also invoked by the challenge of norms. As we have seen norms carry some implicit 'meaning' (or functional significance) for the agent. Cognitivism is based on the use of symbolic representation—some salient characteristic is represented in the mind as a symbol. In cognitivist systems the meaning of the symbol must be provided from outside or coded into the system.

In cognitivist approaches then, the frames the agent can use to judge the salience of a norm as well as any functional significance of that norm must be provided from outside and therefore are not under the control of the system. Such agents can generate

emergent behaviour but not in a manner analogous to the way humans appear to in relation to norms. To simulate the emergence of norms what is significant and meaningful must be allowed to change as a result of the interaction between agents as individual (micro) choices shape social (macro) consequences on coupled viability landscapes.

We must conclude, therefore, that it is difficult to do justice to the emergent nature of norms using cognitivist approaches.

The framing problem presented major problems for even simple robots attempting to navigate their way in relatively fixed environments. The solution was an approach to cognition which allowed agents to learn and evolve their parameters to deal with environments instead of attempting to program in the necessary contingency table. The resulting *connectionist* models (Brooks, 1991) invoke no symbols, thereby avoiding some aspects of the grounding problem. Rather than manipulating symbols which 'stand for something' in the agent's environment, meaning is embodied in fine-grained structure and pattern throughout the network. Connectionist approaches can derive pattern and meaning by mapping a referent situation in many different (and context dependent) ways. Meaning in connectionist models is embodied by the overall state of the system in its context. It is implicit in the overall '*performance in some domain*'. Connectionism led to a major leap forward in robotics. However Dreyfus has identified a residual challenge that confronts both cognitivism and connectionist approaches. This is how to '*directly pick up significance and improve our sensitivity to relevance*' … since this ability '*depends on our responding to what is significant for us*' given the current contextual background (Dreyfus, 2007: 30).

Linking this to thinking about norms, a connectionist model could converge on a pattern within its environment and develop an effective accommodation to it. If that pattern changes in a novel way—one not anticipated by the system designers—a connectionist system may still be able to accommodate that change within limits. What it still cannot do is make a judgement as to how the new pattern is in its interests and nor can it initiate strategies to attempt to influence that new pattern to turn it to its advantage, except to the extent that some representation or implicit design aspect framed from outside (i.e. through the hand and mind of the system designer) specifies where the boundaries of self interest are—it does not have and cannot develop the agency which, as has been discussed above, may play a role in normative action.

Connectionist approaches therefore support experimentation into aspects of norms where there is some scope for habits to form and adjust in relation to changing contexts, including the behaviour of other agents. However, as Froese et al. argue '*as long as there is no meaningful perspective from the point of view of the artificial agent, which would allow it to appropriately pick up relevance according to its situation in an autonomous manner, such a system cannot escape the notorious* "*frame problem*"' (Froese & Ziemke, 2007: 8).

This brings us to the argument for enactive approaches to artificial systems. Enactivism solves the framing and grounding problem in the manner already

described earlier in this chapter—the self-producing nature of the agents provides them with a fundamental goal—maintenance of self.

From what has now been considered throughout this chapter it is now possible to identify the minimum set of requirements for an approach to simulation capable of reproducing dynamics which are reasonable analogues of social norms in human social systems. To simulate norms we need agents:

- Who's state at any given time is a product of its interactions with other agents.
- Have a low-level goal (this presupposes a minimal condition which they seek to maintain such as their viability) and against which their actions and the actions of others can be evaluated.
- The range of emergent norms that will arise in such a system is influenced by the substantive constitutive nature of the agents—and hence the range of states they are capable of recognizing (perceiving), evaluating and responding to.
- This must involve more than a capacity to simply couple to the environment but a capacity to break symmetry (Barandiaran et al., 2009).

Towards an Emergentist Simulation of Norms

We are still a considerable way from being able to build systems with these capabilities. On the positive side we are getting closer to being able to specify what it will take.

1. We need to be able to model an agent as an operationally closed (autonomous) entity. This does not have to be at the level of biological process—the agent does not need to produce itself in a material sense; rather as Froese and Ziemke (2007) state the artificial system must be capable of generating its own systemic identity at some level of description. The level of description will be relative to our purpose for performing the simulation and which aspect of social (including normative) functioning we are attempting to explore.
2. An artificial system must have the capacity to actively regulate its ongoing sensori-motor interaction in relation to a viability constraint linked to the maintenance of its identity.
3. Agents need to be able to be assembled (or to self-assemble) onto coupled fitness landscapes where the fitness function is linked to the underlying viability set.

A recipe for working towards such an artificial system has been sketched by Morse et al. (Morse, Lowe, & Ziemke, 2008) and many simple practical experiments conducted in this direction (see for example Di Paolo n.d.; Di Paolo & Lizuka, 2007; Froese & Di Paolo, 2008; Montebelli, Lowe & Ziemke, 2009).

Patterns which emerge in the relationship between such agents would qualify as norms in that they would be genuinely emergent. They would represent quasi-stable patterns which satisfy the viability requirements of the participating agents. It will,

however, likely be very difficult for a human observer to understand in what way these patterns are 'meaningful' (i.e. functional with respect to the agents and/or the system they comprise) other than in the highly abstract context of the artificial world. It will be difficult to steer the emergence of such patterns towards particular experimental ends as well as to interpret what they suggest by way of outcome. De Loor et al. (n.d.) suggest that one approach to this problem may be to include a real human as a participating agent. No doubt we will discover more of how this might be possible as we progress towards the development of simulation platforms with these types of characteristics. If the slow rate of progress within AI is a guide, this will not be a rapid process.

Conclusion

Norm-following agents are characterised by being able to generate alternative response through their interaction with one another which serve to maintain each as a viable entity within particular social domains. Norms represent quasi-stable patterns or attractors generated by the process of mutual accommodation on coupled viability landscapes. These accommodations arise through multiple modes of interaction from reflex, through affectively modulated interactions through to tactical and strategic positioning made possible by different levels of cognitive capability extending in humans to agency and identity and the scope for language as a particularly flexible mode for mutual influence.

To date attempts to study norms within social science have failed due to the micro–macro divide—the inability by contemporary social science to provide an adequate account of the dialectic between macro social structures and individual dispositions and action. While systems thinking, particularly that associated with complex systems, has significantly advanced our understanding of mechanism of emergence and therefore served to illuminate mechanisms associated with this dialectical interpenetration of levels, it has done relatively little to date to contribute to our understanding of the particular way in which this may operate in human social systems. Nevertheless these advances, as well as rudimentary social simulations, made possible through cognitivist and more recent connectionist approaches to robotics, have proceeded hand in glove to help advance our understanding of human social system dynamics which extend well beyond what was achieved in the past several hundred years within social theory and philosophy alone.

Key and often fresh insights into what a next generation of social simulation platforms might look like can be drawn from recent advances in cognitive biology and evolutionary robotics. Unlike much social simulation, which has tended to stay with representational approaches, these other fields have taken seriously the questions posed by the entanglement of cognitive capabilities as well as the known problems with cognitivist approaches, in particular the framing and the grounding problems. This work has helped us to identify what the characteristics of a system need to be to support investigation of norms.

References

Archer, M. (1998). Realism in the social sciences. In M. Archer, R. Bhaskar, A. Collier, T. Lawson, & A. Norrie (Eds.), *Critical realism: Essential readings*. London: Routledge.

Archer, M., Bhaskar, R., Ciollier, A., Lawson, T., & Norrie, A. (1998). *Critical realism: Essential readings*. London: Routledge.

Ashby, W. R. (1974). Self-regulation and requisite variety. In F. E. Emery (Ed.), *Systems thinking*. Great Britain: Penguin.

Barandiaran, X. (2005). *Behavioral adaptive autonomy. A milestone on the ALife roue to AI?* San Sebastian, Spain: Department of Logic and Philosophy of Science University of the Basque Country.

Barandiaran, X., Di Paolo, E. A., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry and spacio-temporality in action. *Adaptive Behavior, 17*(5), 367–386.

Barandiaran, X., & Moreno, A. (2006). On what makes certain dynamical systems cognitive: A minimally cognitive organization program. *Adaptive Behavior, 14*(2), 171–185.

Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review, 10*(1), 20–46.

Berger, P. L., & Luckman, T. (1972). *The social construction of reality*. Harmondsworth: Penguin.

Bhaskar, R. (1997). *A realist theory of science*. London: Verso.

Bhaskar, R. (1998). *The possibility of naturalism*. London: Routledge.

Brooks, R. A. (1991). Intelligence without representation. *Intelligence Without Reason, 47*, 569–595.

Castelfranchi, C. (1998). Simulating with cognitive agents: The importance of cognitive emergence. In J. S. Sichman, R. Conte, & N. Gilbert (Eds.), *Multi-agent systems and agent based simulation*. Berlin: Springer.

Clark, A. (1998). *Being there: Putting brain, body and world together again*. Cambridge, MA: The MIT Press.

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences, 3*(9), 345–351.

Coleman, J. S. (1994). *Foundations of social theory*. Cambridge, MA: Belknap.

Colombetti, G., & Thompson, E. (2008). The feeling body: An enactive approach to emotion. In W. F. Overton, U. Muller, & J. L. Newman (Eds.), *Developmental perspectives on embodiment and consciousness*. London: Lawrence Erlbaum Associates.

Cosmelli, D., Lachaux, J.-P., & Thompson, E. (2007). Neurodynamics of consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness*. Cambridge: Cambridge University Press.

Cowley, S. J., & Macdorman, K. F. (2006). What baboons, babies and tetris players tell us about interaction: A biosocial view of norm-based social learning. *Connection Science, 18*(4), 363–378.

Crossley, N. (2001). The phenomenological habitus and its construction. *Theory and Society, 30*(1), 81–120.

Damasio, A. (2000). *The feeling of what happens: Body, emotion and the making of consciousness*. London: Vintage Books.

Damasio, A. (2006). *Descartes' error*. London: Vintage Books.

De Loor, P., Manac'h, K., Fronville, A., & Tisseau, J. (2008). Requirements for an enactive machine: Ontogenesis, interaction and human in the loop. Plouzane, France: Universite, Europeenne de Bretagne.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences, 4*(4), 429–452.

Di Paolo, E. A. (n.d.). Organismically-inspired robotics: Homeostatic adaptation and teleology beyond the closed sensorimotor loop. Brighton: School of Cognitive and Computing Sciences, University of Sussex.

Di Paolo, E. A., & Lizuka, H. (2007). How (not) to model autonomous behaviour. *Biosystems, 91*(2), 409–423.

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology, 20*(2), 247–268.

Ekman, P. (1992). Facial expressions of emotion: New findings, new questions. *Psychological Science, 3*(1), 34–38.
Ellis, G. F. R. (2006). On the nature of emergent reality. In P. Clayton & P. Davies (Eds.), *The re-emergence of emergence: The emergentist hypothesis from science to religion*. Oxford: Oxford University Press.
Engels, F. (1934). *Dialectics of nature*. Moscow: Progress Publishers.
Franklin, S. (1998). *Artificial minds*. London: MIT Press.
Froese, T. & Di Paolo, E. A. (2008). Stability of coordination requires mutuality of interaction in a model of embodied agents. Paper presented at the Procedures of the 10th international conference on simulation of adaptive behavior, Berlin, Germany.
Froese, T., & Di Paolo, E. A. (2009). Sociality and the life-mind continuity thesis. *Phenomenology and the Cognitive Sciences, 8*(4), 439–463.
Froese, T., & Ziemke, T. (2007). *Enactive artificial intelligence*. Brighton: University of Sussex.
Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences, 4*(1), 14–21.
Gardenfors, P. (2006). *How homo became sapiens: On the evolution of thinking*. Oxford: Oxford University Press.
Gibbs, J. P. (1965). Norms: The problem of definition and classification. *American Journal of Sociology, 60*, 8.
Gibson, J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing*. New York, NY: Lawrence Erlbaum.
Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Berkeley, CA: University of California Press.
Gilbert, N. (2002). Varieties of emergence. Paper presented at the social agents: Ecology, exchange, and evolution conference, Chicago.
Goldspink, C., & Kay, R. (2007). *Social emergence: Distinguishing reflexive and non-reflexive modes, AAAI fall symposium: Emergent agents and socialites: Social and organizational aspects of intelligence, Washington, DC*. Menlo Park, CA: AAAI.
Goldspink, C. & Kay, R. (2008). Agent cognitive capabilities and orders of emergence: Critical thresholds relevant to the simulation of social behaviours, AISB convention, communication, interaction and social intelligence, April 1–4, 2008, University of Aberdeen.
Hejl, P. (1993). Culture as a network of socially constructed realities. In A. Rigney & D. Fokkema (Eds.), *Cultural participation: Trends since the middle ages* (pp. 227–250). Amsterdam: John Benjamins Publishing Company.
Johnson, M. (1990). *The body in the mind: The bodily basis of meaning, imagination and reason*. Chicago, IL: The University of Chicago Press.
Kauffman, S. A. (1993). *The origins of order: Self organization and selection in evolution*. New York, NY: Oxford University Press.
Kauffman, S. (2000). *Investigations*. New York, NY: Oxford.
Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.
Kravchenko, A. V. (2007). Essential properties of language, or, why language is not a code. *Language Sciences, 29*, 650–671.
Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York, NY: Basic Books.
Lehrer, J. (2009). *The decisive moment: How the brain makes up its mind*. Melbourne, VIC: Text Publishing.
Leont'ev, A. N. (1978). *Activity, consciousness and personality*. Englewood Cliffs, NJ: Prentice Hall.
Lewis, D. (1969). *Convention: A philosophical study*. Cambridge MA: Harvard University Press.
Luhmann, N. (1990). *Essays on self reference*. New York, NY: Columbia University Press.
Luhmann, N. (1995). *Social systems*. Stanford, CA: Stanford University Press.
Maturana, H. (1978). Biology of language: The epistemology of reality. In G. A. Miller & E. Lenneberg (Eds.), *Psychology and biology of language and thought: Essays in honor of Eric Lenneberg*. New York, NY: Academic.

Maturana, H., & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Boston, MA: Reidel.
Menary, R. (2008). Embodied narratives. *Journal of Consciousness Studies, 15*(6), 63–84.
Mingers, J. (1991). The cognitive theories of Maturana and Varela. *Systems Practice, 4*(4), 319–338.
Montebelli, A., Lowe, R., & Ziemke, T. (2009). *The cognitive body: From dynamic modulation to anticipation. Anticipatory behavior in adaptive learning systems* (Vol. 5499). Berlin: Springer.
Moreno, A., & Etxeberria, A. (1995). *Agency in natural and artificial systems*. San Sebastian, Spain: Department of Logic and Philosophy of Science University of the Basque Country.
Morse, A. F., Lowe, R., & Ziemke, T. (2008). Towards an enactive cognitive architecture. Paper presented at the international conference on cognitive systems, April 2–4, 2008, Karlsruhe, Germany.
Ochs, E., & Capps, L. (1996). Narrating the self. *Annual Review of Anthropology, 25*, 19–43.
Reddy, V. (2008). *How infants know minds*. London: Harvard University Press.
Rocha, L. M. (1996). Eigenbehavior and symbols. *Systems Research, 13*(3), 371–384.
Ross, D. (2007). H. sapiens as ecologically special: What does language contribute? *Language Sciences, 29*(5), 710–731.
Rudrauf, D., Lutz, A., Cosmelli, D., Lachaux, J.-P., & Le Van Quyen, M. (2003). From autopoiesis to neurophenomenology: Francisco Varela's exploration of the biophysics of being. *Biological Research, 36*, 27–65.
Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology, 52*, 141–166.
Sawyer, K. R. (2001). Emergence in sociology: Contemporary philosophy of mind and some implications for sociology theory. *American Journal of Sociology, 107*(3), 551–585.
Sawyer, K. R. (2005). *Social emergence: Societies as complex systems*. Cambridge: Cambridge University Press.
Shapiro, L. A. (2011). *Embodied cognition*. Bristol, PA: Taylor & Francis.
Smith, L. B. (2005). Cognition as a dynamic system: Principles from embodiment. *Developmental Review, 25*, 278–298.
Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences, 7*(8), 333–348.
Sripada, C. S., & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers (Ed.), *The innate mind: Culture and cognition* (Vol. 2). New York, NY: Oxford Univer-sity Press.
Stewart, J., Gapenne, O., & Di Paolo, E. A. (2010). *Enaction: Towards a new paradigm for cognitive science*. Cambridge, MA: MIT Press.
Therborn, G. (2002). Back to norms! On the scope and dynamics of norms and normative action. *Current Sociology, 50*(6), 17.
Thompson, E. (2004). Life and mind: From autopoiesis to neurophenomenology, a tribute to Francisco Varela. *Phenomenology and the Cognitive Sciences, 3*, 381–398.
Thompson, E. (2005). Sensorimotor subjectivity and the enactive approach to experience. *Phenomenology and the Cognitive Sciences, 4*(4), 407–427.
Thompson, E., & Varela, F. J. (2001). Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences, 5*(10), 418–425.
van Duijn, M., Keijzer, F., & Franken, D. (2006). Principles of minimal cognition: Casting cognition as sensorimotor coordination. *Adaptive Behavior, 14*(2), 157–170.
van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences, 21*, 615–665.
Varela, F. (1997). Patterns of life: Intertwining identity and cognition. *Brain and Cognition, 34*, 72–87.
Varela, F., Thompson, E., & Rosch, E. (1992). *The embodied mind*. Cambridge, MA: MIT Press.
Vygotsky, L. S. (1962). *Thought and language*. Cambridge, MA: MIT Press.

Chapter 5
It Takes Two to Tango: We-Intentionality and the Dynamics of Social Norms

Corinna Elsenbroich

Wherever I am, there's always Pooh,
There's always Pooh and Me.
Whatever I do, he wants to do,
"Where are you going today?" says Pooh:
"Well, that's very odd 'cos I was too.
Let's go together," says Pooh, says he.
"Let's go together," says Pooh.

Us Two by A. A. Milne

Introduction

Margaret Thatcher's "There is no such thing as society" is one of the defining statements of her premiership, describing a world in which only individuals exist and each and everyone needs to take responsibility for their own actions.[1] The spirit of individualism also pervades the social sciences, starting with microeconomic theory but further invading other social sciences in the form of rational choice, exchange or game theory. It is futile to ask which came first, the individualisation of society or the victory of individualism in the social sciences. They feed back into each other like most social phenomena.

Individualism, however, leaves a problem older than sociology itself: the problem of moral behaviour, of value-lead behaviour, of normative behaviour. The world around us is full of social norms, institutions, pro-social actions; and individualism cannot adequately explain either the variety or the complexity of this social world. This is commonly called the structure/agency problem and it has been a

[1] http://www.margaretthatcher.org/document/106689 2

C. Elsenbroich (✉)
Centre for Research in Social Simulation, University of Surrey, Surrey, UK
e-mail: c.elsenbroich@surrey.ac.uk

M. Xenitidou and B. Edmonds (eds.), *The Complexity of Social Norms*,
Computational Social Sciences, DOI 10.1007/978-3-319-05308-0_5,

long-standing dichotomy of the social sciences; for an excellent review of these positions see Ritzer (2007) or Giddens (1993).

Against the purely individualist position (e.g. Coleman) stands the tradition of structuralism in which society and social structures exist above the individual's behaviour.

> If we cannot be bound by duty except to conscious beings and we have eliminated the individual, there remains as the only other possible object of moral activity the *sui generis* collective being formed by the plurality of individuals associated to form a group … We arrive then at the conclusion that if a morality, or system of obligations and duties, exists, society is a moral being qualitatively different from the individuals it comprises and from the aggregation from which it derives. (Durkheim, 1974, p. 25)

In this quote by Durkheim, society clearly does exist and has the role to be what binds the individual to moral behaviour. Thus society is something over and above the sum of the individuals it comprises and becomes the cause of moral behaviour.

Despite some attempts to bridge the gap between the two extreme positions, like Giddens' Structuration Theory (Giddens, 1984), over time positions involving structures gave way more and more to individualism. The victory of individualism in the social sciences comes, at least in part, from the (perceived) success of neoclassical economics, the study of aggregates of choices of "rational agents", agents that have full information and the goal of personal utility maximisation. Markets are analysed using this research paradigm and for many commodities and goods the theory makes reasonable predictions. The Market is a macro-phenomenon resulting from the exchanges of goods of utility maximising, omniscient agents. Using the idea of a market, theories like exchange theory (Heath, 1976) transported individualist assumptions into sociology and the area of rational choice theory and methodological individualism became well established fields.

As stated before, individualism leaves explanatory gaps, not being able to explain the various and complex social phenomena found in the world. We find a lot more pro-social behaviour and social order than rational choice allows, in particular when looking at smaller aggregates, e.g. families, friendship groups etc. Also, strategic social interaction encoded in game theory falls short of experimental empirical corroboration.

One area of research trying to fill this gap is agent-based modelling, a methodology inspired by the idea of complex systems. Agent-based modelling assumes that society can be modelled as a complex system. The basic idea is that macro-phenomena can be generated from the interactions of simpler parts, in the case of a society, the constitutive agents. Showing macro-phenomena as emergent means ontologically we are committed only to individuals but can still have system properties as emerged phenomena.[2] Individuals with simple specifications interact bringing about the complex social structures we find in society.

[2] For an excellent discussion of emergence in agent-based modelling see Neumann (2006).

Many social phenomena have been modelled using agent-based models and simulations. Models exist at different levels of abstraction, from very abstract models such as Schelling's Segregation model (Schelling, 1971) to models more focused on real-world policy application such as the "Zürich Watergame" model, a participatory model where some of the agents are humans. The differences in the levels of abstraction result from the number of parameters involved in the model and from the complexity of the agents, i.e. whether they have few or many attributes/roles/rules and whether the agents are static or can change/learn/adapt (their cognitive complexity). What agents have in common though is that they are defined strictly individually meaning that they pursue their own goals and social behaviour emerges from the individual behaviours. The ontological commitment resulting from current agent definitions is no higher than that of individualism and we still do some justice to phenomena "over and above" the individual.

Although agent-based modelling adds to strict individualism and the idea of a selfish agent by emergent social phenomena, it does not quite do justice to what could be called the *homo duplex*. The *homo duplex* describes the pull between selfish behaviour and moral or pro-social conduct in human beings. For Durkheim this duality was at the heart of sociological enquiry, in fact, sociology was to be the methodology of a moral science (Giddens 1984). Current agent-based models are able to represent simple normative dynamics, like norm diffusion and adoption but not more complex dynamics such as norm change and norm evolution. In this chapter I look at the possibility of extending our ontology a little in order to model human social or normative behaviour along the lines of the *homo duplex* rather than the *homo economicus* (or a version thereof). What we need is to capture the other side of the human, the side that behaves according to values. Conceptualising the human with this duality enables us to do justice to society as an entity.

In this chapter I explore the concept of we-intentionality or shared intentionality as the foundation of this dual to selfishness. My thesis is that without we-intentionality we cannot explain the complex social world humans inhabit. For this I assume the uniqueness of the human social world (in complexity and abstraction) meaning there must be a unique feature of humans bringing about this unique social world. I will argue that we-intentionality is this unique feature. I will then discuss how we-intentionality might be operationalised in agent-models of normative behaviour.

Section "Introduction" introduces agent-based modelling, a relatively recent methodology using computer modelling to investigate social macro-phenomena. Section "Agent-Based Modelling" discusses the role of intentionality for cultural richness and complexity of social life. Section "Agent-Based Models of Norms" presents two versions of we-intentionality, one reductive and one fundamental. Section "An Impoverished Ontology" discusses how we-intentionality will be of use in agent-based models and finally Sect. "Intention in Agent-Based Models" concludes and points at future research.

Agent-Based Modelling

An agent-based model is a computer program consisting of an environment and a set of autonomous agents, which interact with each other and with the environment. Agents are autonomous in the sense that there is no central decision maker orchestrating behaviour. A simpler form of an agent-based model is a cellular automaton. Cellular automata have been used extensively in the natural sciences to model neighbourhood effects such as the Ising model of the ferromagnet. In a cellular automaton there is no separate environment meaning that the only interactions are between agents (or rather between cells on a grit). In the social sciences cellular automata have been used to reconstruct the emergence of social phenomena such as segregation (Schelling, 1971), the spreading of information (Gilbert & Troitzsch, 2005, pp. 140–142) or the emergence of cooperation among selfish agents (Hegselmann, 2001). For more on cellular automata, see Liebrand, Novak, and Hegselmann (1998) or Gilbert and Troitzsch (2005). Due to their simplicity cellular automata are fairly limited and the modelling ontology has been widened to that of agent-based models.

Agent-Based Models of Norms

Agent-based models have been used for a wide range of phenomena with existing models covering market mechanisms, diffusion mechanisms, class systems, migration patterns, social networks, etc. Agent-based models have also been used for the study of social norms (for an excellent review see Neumann, 2008). In agent-based modelling the study of social norms has so far analysed social norms as emergent features similar to physical or biological structures (Axelrod, 1984; Buchanan, 2007; Epstein, 2000). Individuals are usually defined along the lines of the Belief-Desire-Intention agent (Bratman, 1987). Agents make decisions depending on their beliefs about the environment and other agents (e.g. “food located one step ahead”, “neighbour has food”), their desires (e.g. “hungry”) and their intentions (e.g. “harvest food”, “steal food”). Actions are usually decided on by what is the most beneficial option for the agent. Models of norms most often use punishment to skew agent behaviour towards normative decisions (Axelrod, 1997; Hales, 2002; de Pinninck, Sierra, & Schorlemmer, 2008, etc.) but sometimes embed normative behaviour into the agents’ desires meaning that agents imitate their environment (Epstein, 2000). Either way, the ontology of the agent specification is restricted to individual beliefs, desires and intentions (cf. Bratman, 1987).

These simple agent specifications generate some macro phenomena of normative behaviour, for example the spread of a norm through a population over time, the influence of punishment on large-scale normative adherence, or the function of normative adherence for a population.

There are however, many characteristics of social norms that are not touched upon by these simple models. Neumann (2009) criticises existing models of norms for not touching on important features of normative behaviour derived from research on norms in psychology. One is that normative behaviour is related to emotions such as guilt and shame which none of the existing models capture. Connected to the emotionality of norms is that models of norms do not capture one of the most important features of normative behaviour, namely norm internalisation. Social norms only work because humans are "socialised" over the course of their lives meaning that they internalise social norms as behavioural blueprints. If we model social norms without the aspects of emotions and internalisation, we might be modelling behavioural regularities but we do not model genuine normative behaviour, Neumann argues.

Another critique can be found in Xenitidou and Elsenbroich (2010), this time from the vantage point of social, rather than individual psychology. The authors discuss three kinds of normative behaviour, i.e. conformity, obedience and compliance commonly distinguished in (experimental) social psychology. Agent based models only capture the first two, conformity and obedience, which are relatively simple behavioural mechanisms. Compliance is a norm following behaviour in which an agent actively chooses to adhere to a norm although it might contravene its personal believes and values. It is the normative behaviour touching on the homo duplex, the possibility of the human to forgo an individual advantage for the greater good. None of the present models tackles compliance and with the current agent definitions compliance cannot be modelled as agents do not have the cognitive capacity to reason about behaviour in this more involved way.

And finally, and most pertinent for this book, no agent-based model has modelled the change of norms over time in a society, the dynamic development of norms. While there are always social norms governing society, what counts as normative and deviant behaviour changes over time. Although not exactly normative, in the 1970s it was normal to smoke in many public places such as buses, planes and offices. Overtime the places in which smoking was permitted declined until it was outlawed in most places, even in pubs and bars. From smoking being normal behaviour it has become deviant in most situations. It has variously been argued that many agent-based models are too abstract, too reductive and bear little relation to the real world. Rather than looking for truthfulness modellers have instead been overly concerned with models being parsimonious.

An Impoverished Ontology

A perfect instantiation of this fact is the K.I.S.S. principle (Keep-It-Simple-Stupid) advocating that simpler models are (intrinsically) better models. Arguing against this priority of simplicity in favour of descriptive and truthful models see Edmonds

and Moss (2004). The authors argue that the K.I.S.S. principle is founded on the false premise that simplicity is "truth-indicative", i.e. that there is an inherent reason why simpler models are more likely to be true (for a rebuttal of this premise see Edmonds, 2002). In fact, it is unlikely that a simple model will adequately represent a complex real world phenomenon. The authors advocate the K.I.D.S. principle instead: Keep-It-Descriptive-Stupid. Rather than starting from the simplest possible model (and then adding features if the simple model is inadequate) the starting point should be a model that is as descriptive of given data and evidence as possible. Once this model is understood, it can be simplified if parts are found to be superfluous to the modelling of a specific phenomenon.

There are some good reasons for advocating simplicity. Computer models are difficult to validate and the more complicated the setup, the harder the validation. First of all it becomes more difficult to know whether the model is circular (whether the outcome was programmed into the model) and what the influence of particular parameters on the outcome is. This means if we can reconstruct a phenomenon with less assumptions we should do so; the parsimony requirement of Occam's razor applied to computer simulations. This parsimony does have a downside though. Although parsimony helps to keep control of the model, if the assumptions themselves are too simple or simplistic, the model's adequacy might well suffer. Models of norms are a case in point where it seems that overly simplistic assumptions about the agents lead away from truthfulness. Although one should not assume more than necessary, one should also not assume less than necessary to model a phenomenon. In the case of normative behaviour, this is particularly important as social norms are both so intuitive and elusive.

The visible macro-phenomenon of normative behaviour is a behavioural regularity across a population. However, not every regularity is normative behaviour. We would for example not call the behaviour of the molecules in a ferromagnet "normative" even though they display regular patterns. If we are very inclusive we might classify animal behaviour as normative, for example bird or fish formations, but even if we do so we need to acknowledge that human social norms are considerably more complex than these animal counterparts. Even, regularities in human societies might come about due to reasons other than normativity. For example, the pharmacy Boots is currently giving away free nappy bags to anyone who joins their Parenting Club. As a result, almost every pram is adorned by one of those nappy changing bags. This phenomenon is a behavioural regularity but it does not come about normatively.

Intention in Agent-Based Models

Intention plays a major role in agent specifications in agent-based models. Although some agent systems use very simple agents, as soon as any sort of planning, even at the most rudimentary level, comes into play, agents are modelled on the Belief-Desire-Intention (BDI) framework (cf. Bratman, 1987). Agents have desires or

goals, beliefs about the world and intentions, which are plans towards the goals constructed from the beliefs about the world. In models of normative behaviour, the starting point is usually a BDI agent as simpler agents can almost not be claimed to behave normatively. There is a range of foci for models of normative behaviour. One focus is on cooperation and punishment (e.g. Axelrod, 1997; Hales, 2002; Macy & Sato, 2002). Here the goal of an agent is to maximise individual utility and the intention is to defect when cooperation is detrimental for this goal. Another focus is agents' adaptation to a social environment using imitation or memetics (e.g. Edmonds, 2006; Epstein, 2000; Flentge, Polani, & Uthmann, 2001; Hales, 2001). Here agent behaviour is socially determined, the goal is to "fit in" with others. The intention is to change one's behaviour depending on the social circumstances. A final category of models of norms is to analyse the function of norms for society (Castelfranchi, Conte, & Paolucci, 1998; Conte & Castelfranci, 1995; Saam & Harrer, 1999). The cited models use a utility maximising framework thus having the same goal/intention description as the cooperation models. The dynamics investigated by agent-based models of norms can be roughly classed as emergence and diffusion. As discussed above, two important aspects of normative systems have not been touched on by agent-based models. One is the high complexity and abstraction of human normative social systems. The other is the dynamics of changing norms, e.g. the change from shaking hands to hugging as a greeting, the change from smoking as "cool" to smoking as an outcast activity. What follows is a list of questions about social norms that agent-based models tackled or might want to tackle in future.

1. How can we explain that people behave pro-socially at a cost to them- selves? (In particular without appealing to functional explanations such as "social norms are beneficial for society", even if simulations show that such benefits might indeed exist, cf. Conte and Castelfranci (1995)).
2. How can we explain that social norms stabilise in a society but without necessarily being adopted by all of society? (Or, how can we explain "global diversity and local conformity", cf. Epstein (2000)).
3. How can we explain the complexity and level of abstractness of norms and institutions found in the human world? (Or what is it that distinguishes our institutions from those of other animals, cf. Boyd and Richerson (2005) and Searle (1995)).
4. How can we explain norms developing and changing over time?

The view of social norms underlying the first question is that normative behaviour is costly for individuals and that an individual's behaviour is determined by their personal utility, the *homo economicus*. Punishment is put in place to skew utility functions to make social behaviour pay for the individual. Social norms are restraints on the goals or the intentions of agents. A goal might directly contradict a social norm, e.g. my goal to rid the world of my uncle Ralph by killing him. However, my goal might be fully socially acceptable such as the goal to be a millionaire, a goal shared by a large proportion of the population. Nevertheless, if I intend to become a millionaire by killing my very rich uncle Ralph and inheriting his fortune, social norms will constrain my intentions, leaving my goal intact.

In the second question social norms are just seen as regularities of behaviour, possibly for coordination but possibly without any additional function, e.g. fashions or fads. No cost considerations are taken into account and the search is for a mechanism of diffusion and adoption of a behaviour.

The third question concerns the complexity of norms and culture found in human social life. Although animals display social behaviour and societal structures, the level of human norms and institutions is considerably more complex and abstract and simple imitation mechanisms, in particular of behaviour only, are insufficient to explain this complexity. As discussed above, we are looking for a uniquely human feature that can explain this complexity.

Question 4 addresses dynamics of norms other than emergence or cessation for which it is sufficient to use the concepts of cooperation (with some norm) or defection (against some norm). For normative evolution or norm change we need a population to subscribe to a specific norm and this specific norm to be replaced by or to evolve into another. Subscription to a specific (set of) norm(s) does presuppose a joining of intentions, just like the joint subscription to the rules of grammar does.

At present, models of norms model human societal norms in the same way as we would model animal behaviour, thus leaving the higher complexity found in human societies unexplained. In what follows I argue that the concept of we-intentionality will be essential to the modelling of human normative behaviour as it lies at the heart of actual human sociality. The price we pay in parsimony, I argue, is worth paying for the increase in truthfulness.

Unique Feature or Emerged Phenomenon?

When it comes to explaining the cultural and social complexity of the human world there are two kinds of standpoints: (a) the complexity is simply emerged from the basic properties that also explain animal behaviour (see above) or (b) humans have a unique feature which explains why their social world is different from that of animals. I opt here for the second standpoint. The reason for this is that much energy has gone into the first and not much headway has been made when it comes to the more complex phenomena, whether it is complex institutions, feedback between norms and individuals or the change of norms over time.

I stated above that we are looking for a unique human capacity to explain the social and cultural complexity we find in the world. Here are some hypotheses:

1. *Theory of Mind*: Having a theory of mind means to be able to recognise another individual as having an independent mind with beliefs, desires and intentions, to recognise it as an *intentional agent*. Having a theory of mind means that human beings not only imitate actions (imitation well known from the animal kingdom) but they can imitate intentions, making learning faster and engendering humans with the capacity to make plans and actions more efficient. The main counterargument against the theory of mind as the fundamental feature is that some primates seem to have a rather complex theory of mind but nowhere near the

same level of social complexity (Tomasello, Carpenter, Call, Behne, & Moll, 2005, p. 708).

2. *Language*: Language allows the dealing with complex concepts (e.g. "justice", "marriage", "money") as well as a-spatial and a-temporal planning ("let's meet tomorrow in Rome"). By being able to form these abstract concepts and plans human societies generate the complex structures actually found. The main counterargument against language as the fundamental reason for added complexity is that language itself needs an underlying psychological capacity to make shared reference to abstract concepts possible (Plotkin, 2003, p. 292). Thus, although necessary as the means of communication for the sharing of goals, plans and intentions, it cannot be fundamental.
3. *Mirror-Neurons and Empathy*: A recent finding of Mirror Neurons as the neurological foundation of imitation has lead to a host of hypotheses regarding social learning and understanding. In particular they are seen as the foundation of empathy, the ability to put oneself into "somebody else's shoes". It is similar to the theory of mind but concerns emotions rather than just beliefs, desires and intentions.
4. *We-intentionality*: The idea of we-intentionality is that humans have the capacity to not only recognise another agent's intention but that humans can join intentions thus making cooperation not an accidental byproduct of behaviour or an aberration of behaviour, forced by the threat of punishment. Rather, humans are hard wired to cooperate, sharing goals and making plans together to achieve those common goals (Tomasello, 2009).

The thesis is not that language, theory of mind or empathy are superfluous in the explanation for the complexity of human social life but that they are not sufficient. As Plotkin (2003) states:

> Chimpanzee culture is indeed the sharing of simple motor acts, however this sharing is achieved. Human culture, by contrast, involves sharing knowledge of what a shop is and how it differs from a prison, and of sharing concepts like justice and national pride. You simply cannot arrive at an understanding of the concept of justice, which in part defines a particular culture, by way of the imitation of simple motor acts. (p. 289)

We-intentionality can be seen as the fundamental capacity of the human being leading to language in the form of wanting to share meaning. A theory of mind and empathy might be necessary to find a suitable subject for intention sharing (I cannot share my intentions with my computer) but they are not sufficient to explain the complexity.

Intentionality

In this section I discuss different layers of intentionality. Figure 5.1, taken from Searle (1995), will help guide us through the different kinds and levels of intentionality. The first level is the mire of intentional facts. These can be split into three kinds of intentionality. Individual intentionality is the most basic form where an

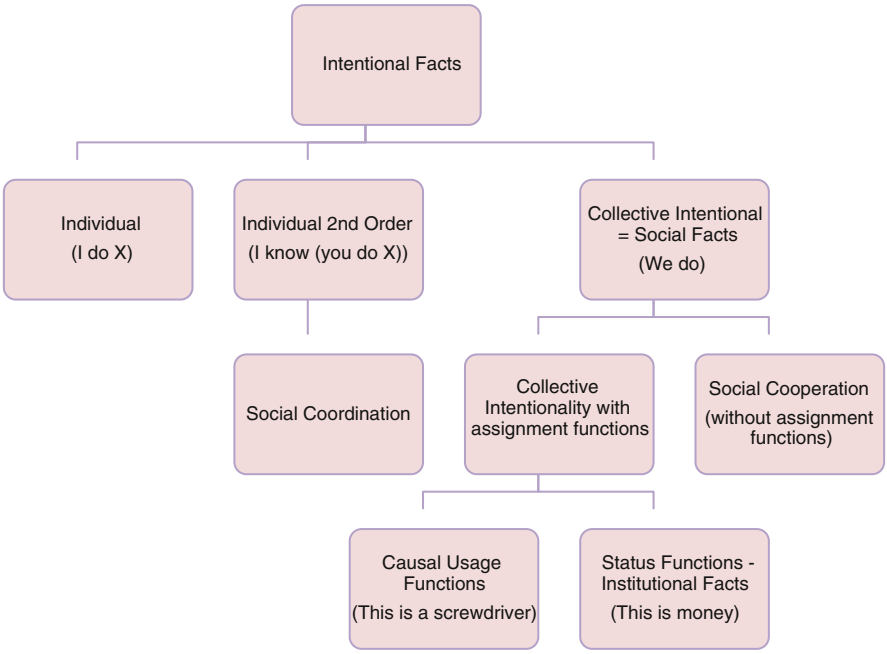

Fig. 5.1 Layers of intentionality

individual's actions are directed towards a goal and action steps are planned towards achieving this goal. This kind of intentionality is operationalised in the BDI agents discussed below.

An individual displays second order intentionality (above "Theory of Mind") if it recognises other individuals as intentional agents, i.e. agents whose actions are directed towards a goal. Second order intentionality has been explicitly operationalised in cognitive agent models such as Bosse, Memon, and Treur (2007). Non-explicitly it lies at the heart of many simulations however. Any game theoretic interaction presupposes the conceptualisation of the interaction partner as behaving intentionally, e.g. Axelrod (1986). Simulations like Lorscheid and Troitzsch (2009) and Andrighetto, Campenni, and Conte (2010) that model social interactions explicitly (e.g. via norm invocation) have an underlying assumption of a theory of mind. Second order intentionality accounts for social coordination.

The final kind of intentionality to be distinguished is collective intentionality, the we-intentionality we want to discuss later on in this chapter. The hypothesis is that collective intentionality lies at the heart of human social behaviour. It is the underlying capacity that leads to proper social cooperation (in contrast to the social coordination resulting from the theory of mind) and to the assignment of "usage/status functions" (e.g. this is a screwdriver, this piece of wood is the queen in a game of chess, this piece of paper is money).

As stated above we are looking for a distinguishing feature of human beings explaining the complexity of the social world. Individual intentionality cannot be a contender for this role; many animals act intentionally, e.g. the cat hunts mice.

A theory of mind develops early in human ontogeny, within the first 12 months of live (usually around 8–12 months). Recent empirical research has shown, however, that attribution of intention is not a uniquely human capacity. Primates also show understanding of the intentions and perceptions of other agents, without, however, engaging in the full social and cultural life of human beings (Tomasello et al., 2005).

The third and final kind of intentionality is called "collective intentionality" here. It is the idea that agents not only have an intention and know about another's intentions but agents can join intentions. This intentionality lies at the heart of cultural and complex human social life for Searle (1995) due to it allowing for the assignment of status functions to objects/events thus leading to the possibility of shared meaning about the social world. Money is only money as long as a buyer and a seller acknowledge it as such and a wedding is only a wedding if the function of "wedding" is associated with the ceremony by all concerned.

We-Intentionality

We-intentionality has been investigated in various disciplines, in particular in psychology (e.g. Plotkin, 2003; Tomasello, 2009; Tomasello et al., 2005), philosophy of psychology (e.g. Dennett, 1987; Searle, 1995), computer science and AI (Bratman, 2006, 2009) and philosophy of sociality (e.g. Gilbert, 2009; Tuomela, 2007). The basic idea is that human beings do not only behave following their own intentions but rather that human beings are unique in joining intentions with other agents.

In the little excerpt from A.A. Milne's *Us Two*, different kinds of intentionality are captured and it can be read either as a reductive instantiation of we-intentionality or a non-reductive one. First Pooh asks what you are doing today, an enquiry into another's intention. Pooh then states that this is "odd" as he had the same intention. This is not a we-intention but two identical yet separate individual intentions. Lastly, Pooh concludes "Let's go together", thus making the coordinated but separate individual intentions into a we-intention. This is the reductive reading. Now assume that Pooh's initial intention is to do something together, i.e. a non-specified we-intention. He then asks what you are doing, proclaims that that was just what he had in mind and suggests to do it together. This way we-intentionality is the foundation of Pooh's inquiry.

In what follows I discuss reductive and non-reductive accounts of we-intentionality and assess their usefulness for agent-based models. I start with the experimental evidence for we-intentionality. I focus on the ontogeny work by Tomasello et al. (2005) on attention and intention sharing. Their research is the best we currently have to strongly suggest that we-intentionality is a fundamental feature

of human cognition. I then discuss a reductive position, i.e. a position that acknowledges that we-intentionality is a feature of human sociality but contends that it can be reduced to individual intentions. A reductive position is presented in Bratman (2006, 2009). Also early work by Tuomela and Miller (1988) is often seen as a reductive account but later on Tuomela (2007) distances himself from reductionism and develops a non-reductive account based on a layering of intentions. I finish in the field of philosophy of sociality and discuss Searle's position on shared intentionality in *The Construction of Social Reality* (Searle, 1995). Searle's position is severely criticised for proposing shared intentionality as a fundamental human capacity without developing a principled account of it (Zaibert, 2003). Zaibert's criticism is correct, i.e. Searle does not provide a principled or explicit account of we-intentionality. The reason for it is that Searle discusses we-intentionality, the disposition of humans to join intentions. An explicit account of it would lead to the discussion of we-intentions, the instantiation of we-intentionality. This however is not Searle's goal. He wants to discuss the importance of the disposition for the emergence of complex institutions and culture we observe. Gilbert (2009) provides a principled account of non-reductive we-intentionality.

Experimental We-Intentionality

Tomasello et al. (2005) have been investigating shared attention and intention for almost two decades. They contend that there is a unique human feature they call "we-intentionality" that none of our primate cousins has. We share a large part of our cognitive skills with our primate cousins, such as the folk psychological specification of human action captured by the BDI agent. Chimpanzees as well as autistic children also show a relatively developed theory of mind, recognising intentions in others more than previously assumed (Tomasello et al., 2005). According to the experiments of Tomasello et al. there is something in addition to recognising intentions in others. This "something" is also recognising other agents' understanding ones own intentions, resulting in "shared attention" (intentional perception) as well as homing in to other agent's intentions, resulting in "shared intention". This triadic interaction of me-object you starts from about one year of age.

Tomasello et al. show experimentally that children around their first birthday can distinguish intentional from non-intentional actions thus experimentally establishing that 1 year olds have a theory of mind. Chimpanzees also have a theory of mind. For example they understand when experimenters point them towards food and react by focusing their attention and moving there. Chimpanzees can also be trained to point and thus share attention, with an experimenter. They do not, however, use pointing either in a natural habitat or between Chimpanzees. This means that although the understanding of other agents as intentional is present, the sharing of attention or intention is not.

Another experiment is on the reaction of toddlers towards intentional and non-intentional action. If an experimenter intentionally drops an object, the toddlers are

not inclined to help picking the object up whereas if the drop is accidental, the toddlers help significantly more often. These experiments do not only show that toddlers have a theory of mind but also show a ready capacity to help others. Toddlers do not only recognise (the) others' intentions but start to participate in the intentions. This shared intention can be seen as social cooperation, distinct from social coordination present in other animal species (see Tomasello & Rakoczy, 2007).

Piaget (REF) argued that pretence play starts as an individual past-time before becoming a social interaction. In contrast, Hannes Rakoczy argues that pretence play is from the start social. Rakoczy investigates the connection of we-intentionality with pretence play, concluding that only with the shared usage functions and status functions, resulting from we-intentionality, can we make sense of pretence play.

These experiments show that we-intentionality is a capacity of humans. They also show that they are not a capacity of other primates.

Reductive We-Intention

As the father of the Belief-Desire-Intention agent it might not be surprising that Bratman also developed a theory of shared intention. Neither is it a surprise that he starts from individual intention to build up his theory. At the root of all social behaviour lies the individual intention, where an individual intention is defined as a plan of action to achieve a goal. Shared intention means more than one individuals have the same plan of action to achieve a (common) goal. However, just having the same intention is not sufficient for having a shared intention. We might both have the intention to paint a house and both set off armed with a paintbrush and roller. The first option is that we both paint all walls, i.e. double paint everything. This is not only not a shared intention but not even a coordinated one. The second option is that we coordinate, say by observing what the other is doing and at least not double paint or bump into each other. This is coordinated action and we might say that we shared the experience, but it certainly does not result from shared intentions, as neither knew about the other's intention at the start. What is missing from the scenario is some cognitive state that makes an action result from shared intentions rather than coordinated or even un-coordinated individual intentions. Bratman gives several other examples of large-scale coordinated behaviour, such as many people opening their umbrellas on the street when it starts raining. He also cites the applause after a concert but rather than seeing it as simple coordinated behaviour reacting to a common stimulus (like the opening of umbrellas on the onset of rain), he classifies it as shared behaviour.

> Again, given a sufficient common understanding of the circumstances, an audience at a Yo Yo Man cello recital may more or less spontaneously arrive at a shared intention to applaud together at the end of the performance as they all recognise to be and recognise that they all recognise to be wonderful. When they applaud they do not merely each individually applaud at the same time. Rather they intentionally applaud together. (Bratman, 2006, p. 7)

Bratman's theory is derived from the following six axioms and a Dependency Principle (DEP).

Definition 5.1

Agents A and B have a shared intention to J if

1. intentions on the part of each in favour of activity J
2. agent A knows that agent B has the intention to J (and vice versa)
3. both have intentions in favour of meshing subplans to achieve J
4. beliefs about the joint efficacy of the relevant intentions
5. beliefs about interpersonal intention-interdependence

(DEP) agent A continues to intent to J if and only if agent B continues to intent to J (and vice versa)

6. common knowledge of 1–6 and (DEP).

According to this position, shared intentions derive from individual intentions plus some connecting conditions linking intentions between agents. These conditions are about knowing of the other's intention, agreeing to find ways to achieve a common goal and both seeing the possibility to achieve joining intentions, i.e. plans to achieve the common goal. In addition, there are some conditions on persistence and continuation and the public knowledge of the connectedness of the intentions. This account relies on the assumption of an independent standpoint available to judge whether the conditions are met.[3] Other reductive accounts can be found for example in Kutz (2000) and Zaibert (2003).

Non-Reductive We-Intention and We-Intentionality

Recent literature, however, argues more and more for non-reductive accounts of shared intentionality, see for example Roth (2004), Tuomela (2007) or Schmid (2008). On the non-reductive side we discuss Gilbert as an explicit non-reductive account of we-intention and Searle as a non-reductive account of we-intentionality.

Gilbert's Plural Subject Account of We-Intention

Gilbert (2009) develops an explicit, non-reductive account of we-intention. Against Bratman she develops a "plural subject account" of shared intentions. In short, shared intentions create a *plural subject*, i.e. a body consisting of those agents sharing

[3] Thanks to an anonymous referee for this comment.

intentions, rather than the purely relational account of joint individual intentions by Bratman. Gilbert's plural subject account has parallels to Durkheim's idea of society as the cause of moral behaviour. Gilbert argues that her account more faithfully represents our intuitive understanding of we-intention. The main points are connected to joint commitment and breach of commitment. Agents that have joint intentions have also joint sub-plans and reneging on a sub-plan constitutes a breach of commitment. Gilbert states criteria for we-intentionality the plural subject account satisfies while the joint individual intention account by Bratman does not.

Gilbert identifies three criteria to assess whether a situation is the result of we-intentions. The disjunction criterion states that

> an adequate account of shared intention is such that it is not necessarily the case that for every shared intention, on that account, there be correlative personal intentions of the individual parties. (p. 172)

This criterion supports a non-reductive approach to we-intention as a shared intention does not need an individual intention underlying it. The second criterion is the concurrence criterion stating that

> an adequate account of shared intention will entail that, absent special background understandings, the concurrence of all parties is required in order that a given shared intention be changed or rescinded, or that a given party be released from participating in it. (p. 173)

Once a shared intention is initialised, any changes to the intention have to be agreed by all parties involved. The final criterion, the obligation criterion states that

> an adequate account of shared intention will entail that each party to a shared intention is obligated to each to act as appropriate to the shared intention in conjunction with the rest. (p. 175)

The obligation criterion covers the distribution of rights amongst the parties involved in a shared intention. Reneging on a shared intention breaches the contract made by the shared intention and the party reneging can be rebuked or punished by the other(s). Gilbert shows that an account of we-intention as correlative individual intentions does not satisfy any of these three criteria and develops instead the plural subject account.

> Members of some population P share an intention to do A if and only if they are jointly committed to intend as a body to do A. (p. 179)

Two terms need explanation here, *joint commitment* and intend *as a body*. The joint commitment is constituted by all parties sharing an intention and openly expressing to do so.

> In the basic case, on which I focus here, each of two or more people must openly express his personal readiness jointly with the others to commit them all in a certain way. (page no.)

The expression "as a body" means that the shared intentions will lead to all parties participating in the same instance of the intention. Gilbert's example case is "Sally and Tim [being] jointly committed to intend as a body to produce, by virtue of the actions of each, a single instance of going for a walk with the two of them as the participants in that walk." (p. 181)

Table 5.1 The difference between We-intention and We-intentionality

	Non-reductive	Reductive
Explicit	We-intention as something over and above the linking of individual intentions (e.g. Gilbert, later Tuomela)	We-intention as an explicit add on to or a contract between individual intentions, (e.g. Bratman, early Tuomela)
Implicit	We-intentionality as a disposition of human beings (and other animals) (Searle, Tomasello)	

Gilbert's presents a principled account for a non-reductive version of we-intentions. She develops an account in which the joining of intentions creates a new entity, the plural subject, in which the individuals partake with their actions and to which the individuals are committed in a more involved way of commitment than Bratman's reductive account.

Although this account is non-reductive in the sense of we-intentions being something over and above the set of individual intentions, it might be accused of being reductionist in the sense that it does not allow for the emergence of we-intentions.[4] As every agent has to actively partake in the joint intention there is no space for partially overlapping intentions or the partial sharing of meaning. Most of our social life is not of this kind of commitment, however. We do not actively agree that a particular object is a hammer. We do not actively agree with each other that paper issued by the Bank of England is money. We do not actively agree with each other that we do not bump into each other in the street even though we expect that most people will not do so. We have to distinguish between an account of we-intentionality as an underlying disposition of complex social interaction and an account of the instantiation of we-intentions (see Table 5.1).

Searle's Collective Intentionality

Searle calls the extension of his account of individual intentionality to a system with multiple agents sharing intentions an account of "collective" intentionality (see above Fig. 5.1). Humans exercise collective intentionality by assigning functions to objects or events leading to constitutive rules of action. The pretence play discussed above exemplifies that understanding of status functions of objects, the shared meaning of status functions and the ability to playfully forego the functions. Other examples are the assignment of the function of a hammer to a heavy piece of metal (or other hard material) attached to the top of a stick and the assignment of currency to a piece of paper issued by the Bank of England to be used in market transactions. The first is the assignment of a usage function, i.e. this object is for hammering nails into the wall, the second the assignment of a status function, i.e. this object is money.

[4] Thanks to an anonymous referee for pointing this out.

Searle's view on collective intentionality has its origins in his social philosophy. He demarcates beliefs and desires from intentions in that beliefs and desires share links to the outside world whereas intentions are purely self-referential. I can believe both that I will become a philosopher and you become a philosopher. I can also desire both that I will become a philosopher and you will become a philosopher. I can intend to become a philosopher but it makes no sense for me to intend for you to become a philosopher.

> I cannot intend that my wife is happy, but I of course could intend to make her happy. (Zaibert, 2003, p. 58)

Beliefs, desires and intentions also differ in their satisfaction criteria.

> For example, let us suppose that Jack intends to kill his neighbor Jill. He has been planning to kill her for a while. One day he goes to a store to buy a weapon. While he is driving to the store, a careless pedestrian walks right in front of Jack's car, and he tries in vain to avoid the collision. The pedestrian dies instantly. Suppose that the pedestrian happened to be Jill. (Zaibert, 2003, p. 58)

Jill's death satisfies the sentence "Jack desired to kill Jill" but it does not satisfy "Jack intended to kill Jill", even though he killed her. For intent to be satisfied, there must be a causal chain from the intention to the act or action. (Searle uses these interchangeably).

Searle acknowledges that many people avoid a non-reductive account of we-intentionality as it seems to commit us to an ontology of almost Hegelian proportions in which a "collective spirit" is housed along with individuals (Searle, 1995, p. 25). Both Durkheim's account of morality and Gilbert's account of we-intention smack of this distended ontology. For Searle there is, however, no contradiction between we-intentionality and methodological individualism. Clearly, mental states only exist in individuals' minds; however, one kind of individual mental state is a we-intention. For Searle collective intentionality is the distinguishing feature of human beings for the establishment of institutional facts or complex social norms. He distinguishes observer-relative facts (e.g. "The moon looks beautiful tonight") from brute facts (e.g. "The moon causes the tides"). The moon will still cause the tides if no one is looking, however the first sentence makes no sense without an observer. From this distinction Searle invokes collective intentionality to get from observer relative to institutional facts. Through collective intentionality we can collectively assign a function to an object (e.g. the function of value storage and exchange currency for a piece of paper) and set a set of constitutive rules that make the object fulfil the function (e.g. to be issued by the Bank of England). In a way, collective intentionality is the enabler for shared meaning, as Searle explicitly acknowledges that language is the underlying construction of a social fact. The capacity for we-intentionality is however, prelinguistic as Tomasello's (ref) work shows.

Zaibert (2003) strongly criticises Searle's account of collective intentionality. His main criticisms are that Searle does not develop a principled account of collective intentionality and that collective intentionality is inconsistent with Searle's original account on individual intentionality (Searle, 1983). This inconsistency results from Searle making a distinction between intention and intentionality, with intention being only one instantiation of intentionality in the individual account but in the

collective account not adhering to this distinction. As a result, Zaibert argues, Searle creates an account of collective intentionality but not of collective intention.

Zaibert's criticism is correct, in that Searle does not provide an explicit account of we-intention. Searle's account is we-intentionality (in contrast to we-intention) as the disposition enabling humans to share meaning and through it create institutional facts. For an account of emergent social phenomena we need an implicit account of we-intentionality as an explicit account will only account for instantiations, for the we-intentions. Explicit accounts do not allow for partially shared intentions and the emergence of institutions.

We-Intentionality in Agent-Based Models

In this Section I briefly relate the different conceptions of we-intentionality to agent-based models. This discussion is not an implementation of we-intentionality into agent-based models but rather meant to prepare the ground for modellers to start implementing we-intentionality.

For a reducible notion of we-intentionality we can keep the existing BDI architecture. Looking at Bratman's account the important features of we- intentionality are:

1. The agents agreeing on a shared goal
2. The merging of sub-plans towards the achievement of the goal
3. The common committing to the goal and plans
4. Both stop the activity when one drops out

Bratman's account is an explicit account of we-intentions. Agents form individual goals and intentions and join them up in a sort of contract. Goals are simple statements like "spend a weekend in New York", "paint the house" or "getting married". In an implementation agents can choose common goals out of a set of possibilities or the intersection of their respective goals and agree on one of them. Common commitment can also be modelled with relative ease, even though it will not be descriptive of what happens in real life. Let common commitment be given as a binding contract agents "sign". Defaulting on the contract might lead to punishment.[5] The BDI agent has the sharing of goals and the (contractual) commitment to sub-plans added to its behavioural repertoire.

[5] Note that although the implementation includes implicit punishment, the punishment is only at the point at which there was first a common agreement to do something together. This relates to Gilbert's obligation criterion of shared intentionality (Gilbert 2009). Punishment linked to commitment is related to the literature on theories of fairness intentions; cf. Gintis, Bowles, and Boyd (2006). Punishment linked to commitment can be made dependent on several variables of the actual situation, such as how much the shared goal depends on the sharing, how early on in the process defection occurs, on past behaviour or whether any doubt was raised before setting off in pursuit of the common goal (cf. Gilbert's concurrence criterion). Differences in punishment for different defections would be in line with the findings of Falk, Fehr, and Fischbacher (2008), which show that punishment is both outcome orientated (dependency of shared goal on cooperation) and dependent on the attribution of intentionality to the defecting agent (e.g. repeat defection).

As the reductive we-intentionality extension is a behavioural extension of the BDI agent, basically any model using this architecture can be extended by adding the above behaviour components. The main problem with the reductive account of we-intentionality is that it adds normative behaviour (in the form of cooperation contracts) explicitly to a model. It is thus not possible to produce a model of emergence using this explicit account of we-intention.

But just looking at we-intentionality as a non-reducible will not lead to an emergentist account of normative behaviour. If we take Gilbert's (2009) non-reductive plural subject account of we-intentions, we still end up with agents explicitly bound to explicit we-intentions. There is no space, as stated before, for the non-explicit joining into common meaning, i.e. usage/status functions and social cooperation, e.g. bumping into each other in the street.

An emergentist account can only be achieved by implementing we-intentionality as an agent disposition. This means we need to not only make the agents *behave* according to shared intentions but the agent architecture itself has to incorporate the capacity for shared intentions. Searle assumes we-intentionality to be fundamental and non-reducible to individual intentionality. Looking at the findings of Tomasello and Rakoczy (2003); Tomasello et al. (2005); Tomasello (2009), there is evidence that we-intentionality is indeed a basic human capacity. There are two main features of we-intentionality. One is the recognition of the other as an intentional agent, the theory of mind (cf. Plotkin, 2003). Not only is an agent aware of its own intentions but also has a theory of another agent's intentions. In addition to the recognition of intentions in other agents we need the sharing of intentions, as for example expressed in the helping behaviour of toddlers described above and the sharing of attention, i.e. believing the other agent to understand my intentions in return. This will lead to a partial (or complete) behavioural consensus (note, not necessarily explicitly) and a shared conceptual space (in the form of meaning and usage/status functions).

Models of concepts we might want to use as groundwork are models of language evolution for the emergence and evolution of shared meaning. For example Hutchins and Hazlehurst (1995) develop a simulation of the emergence of a shared lexicon. The simulation replicates the dynamics between external objects/situations, the internal cognitive state of an agent and the communicative social actions. The authors define a lexicon as "a consensus on a set of distinctions" (Hutchins & Hazlehurst, 1995, p. 6). The results show the emergence of just such a shared lexicon.

Building on the emergence of shared meaning would capture Searl's theory of the social construction of reality in which meaning is generated by collective ascription of function to an entity or situation, thus exemplifying shared intentionality. Simulations of language emergence or the emergence of shared meaning can also be directly related to the simulation of we-intentionality as their starting point is we-intentionality in the assumption that agents will communicate with each other, similar to the joint attention discussed in Tomasello et al. (2005). A language is a structure of shared meanings. Social norms can be seen as shared meanings established in a similar way to a lexicon. This would make social norms the lexicon of social structure, just like a language lexicon is of natural structures (Hutchins & Hazlehurst, 1995).

Also some models of normative behaviour could be adapted to the we-intentionality hypothesis. One of the most sophisticated models of normative behaviour currently on the market is the model of norm invocations developed in Lorscheid and Troitzsch (2009). In this model, agents have the goal to colour the world a specific colour (some agents have blue and some have red in the original simulation). They move about a grid colouring the patches they stand on. In addition to colouring they can send norm invocation messages to other agents. For example, if a blue agent sees an agent painting patches red it can send a message asking it to change its colouring to blue. In the original model agents change their colouring behaviour gradually towards the invocated colour. In this scenario it is a possibility to replace simple norm invocation by shared intentionality, i.e. agents sharing a colouring intention.

Clearly, adding we-intentionality to the agent architecture is less parsimonious complicating the simple BDI architecture. The standard BDI architecture is, however, not able to adequately model many dynamics of normative behaviour, such as the feedback between social norms and individual behaviour or the change of norms over time. Although parsimonious, the BDI architecture is insufficient for the modelling of all but the most basic dynamics of normative behaviour. More recent architectures are the agent architecture called EmiL-A (Andrighetto & Campenni, 2007) or EmiL-I-A (Andrighetto, Villatoro, & Conte, 2009). EmiL-A is an architecture which has a normative reasoning component added to a simple BDI architecture. The agents are now not only able to reason with factual beliefs. The normative board contains normative beliefs and normative goals leading to normative action plans. EmiL-I-A is a further extension with an internalisation component. Through the internalisation component the agent can learn new normative beliefs which are then incorporated into its normative reasoning.

It seems any model wanting to go beyond the most basic patterns of social norms and model human normative behaviour needs to extend its set of assumptions beyond the basic BDI architecture, thus becoming less parsimonious. In the face of this the added support to the concept of we- intentionality from the research by Tomasello et al. seems preferable to purely pragmatic additions.

Discussion and Conclusion

I have discussed the main positions on we-intention and we-intentionality, explicit, reductive, explicitly non-reductive and dispositional, and related it to agent-based modelling of social norms. An implementation of we-intentionality leads to less parsimonious agent-based models. Either we-intentionality is added as a behavioural component to a purely individualistic agent (Bratman) or we-intentionality is added as a disposition to the agent architecture (Searle, Tomasello). I argued that any model that wants to go over and beyond simulating the most basic patterns of normative behaviour (e.g. diffusion) need to go beyond a simple BDI agent.

In particular, adding we-intentionality to our agent specification enables us to tackle questions that cannot be tackled without the concept of we-intentionality, such as to model the dynamic change of social norms or the fact that human normative systems are very complex. The non-reductive account of we-intentionality also gains validity from the ontogenetic work of Tomasello et al. which strongly suggests we-intentionality as a fundamental property of human beings. Future work is to look at actual implementation possibilities of we-intentionality into agent-based models.

References

Andrighetto, G., & Campenni, M. (2007). On the immergence of norms: A normative agent architecture. In AAAI Symposium, Social and Organizational Aspects of Intelligence.

Andrighetto, G., Campenni, M., & Conte, R. (2010). Making the theory explicit: The EMIL-A architecture. In EMergence in the loop: Simulating the two way dynamics of norm innovation, chapter 9, pp. 77–88.

Andrighetto, G., Villatoro, D., & Conte, R. (2009). Norm internalisation in artificial societies. *AI Comm European Workshop on Multi-Agent Systems (EUMAS), 23*(4), 325–339.

Axelrod, R. (1984). *The evolution of cooperation*. New York, NY: Basic Books.

Axelrod, R. (1986) An evolutionary approach to norms. *American Political Science Review, 80*, 1095–1111.

Axelrod, R. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton, NJ: Princeton University Press.

Bosse, T., Memon, Z., & Treur, J. (2007). A two-level BDI-agent model for theory of mind and its use in social manipulation. In *AISB 2007 Workshop on Mindful Environments*.

Boyd, R., & Richerson, P. J. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago, IL: University of Chicago Press.

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

Bratman, M. E. (2006). Dynamics of sociality. *Midwest Studies in Philosophy, 30*(1), 1–15.

Bratman, M. E. (2009). Shared agency. In C. Mantzavinos (Ed.), *Philosophy of the social sciences: Philosophical theory and scientific practice*. Cambridge: Cambridge University Press.

Buchanan, M. (2007). *The social atom*. New York: Bloomsbury.

Castelfranchi, C., Conte, R., & Paolucci, M. (1998). Normative reputation and the cost of compliance. *Journal of Artificial Societies and Social Simulation, 1*(3).

Conte, R., & Castelfranci, C. (1995). Understanding the functions of norms in social groups through simulation. In R. Conte & N. Gilbert (Eds.), *Artificial societies. The computer simulation of social life*. London: UCL Press.

de Pinninck, A., Sierra, C., & Schorlemmer, M. (2008). Distributed norm enforcement via ostracism. *Lecture Notes in Computer Science, 4870*, 301–315.

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

Durkheim, E. (1974). *Sociology and philosophy*. New York: Simon & Schuster.

Edmonds, B. (2002). *Simplicity is not truth-indicative*. In Gershenson, C. et al. (2007) Philosophy and Complexity. World Scientific, 65–80.

Edmonds, B. (2006). The emergence of symbiotic groups resulting from skill-differentiation and tags. *Journal of Artificial Societies and Social Simulation, 9*(1), 10.

Edmonds, B., & Moss, S. (2004). From kiss to kids: An 'anti-simplistic' modelling approach. In *Multi agent based simulation* (pp. 130–144). Berlin: Springer.

Epstein, J. (2000). *Learning to be thoughtless: Social norms and individual com-puting*. Technical report, Center on Social and Economic Dynamics Working Paper, No. 6.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness—intentions matter. *Games and Economic Behavior, 62*, 287–303.

Flentge, F., Polani, D., & Uthmann, T. (2001). Modelling the emergence of possession norms using memes. *Journal of Artificial Societies and Social Simulation, 4*(4).

Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Los Angeles, CA: University of California Press.

Giddens, A. (1993). *New rules of sociological method*. Cambridge: Polity Press.

Gilbert, M. (2009). Shared intention and personal intentions. *Philosophical Studies, 144*, 167–187.

Gilbert, N., & Troitzsch, K. G. (2005). *Simulation for the social scientist* (2nd ed.). Berkshire: Open University Press.

Gintis, H., Bowles, S., & Boyd, R. T. (Eds.). (2006). *Moral sentiments and material interests: The foundations of cooperation in economic life*. Cambridge, MA: MIT Press.

Hales, D. (2001). *Tag based co-operation in artificial societies*. Ph.D. thesis, Department of Computer Science, University of Essex.

Hales, D. (2002). Group reputation supports beneficent norms. *Journal of Artificial Societies and Social Simulation, 5*(4).

Heath, A. (1976). *Rational choice and social exchange: A critique of exchange theory*. New York: Cambridge University Press.

Hegselmann, R. (2001). Verstehen sozialer strukturbildungen. In M. Wink (Ed.), *Vererbung und milieu* (pp. 355–379). Heidelberg: Springer.

Hutchins, E., & Hazlehurst, B. (1995). How to invent a lexicon: The development of shared symbols in interaction. In N. Gilbert & R. Conte (Eds.), *Artificial societies: The computer simulation of social life*. London: UCL Press.

Kutz, C. (2000). Acting together. *Philosophy and Phenorerzologica Rlesearch, LXI*(1), 1–31.

Liebrand, W., Novak, A., & Hegselmann, R. (Eds.). (1998). *Computer modelling of social processes*. London: Sage Publications.

Lorscheid, I. & Troitzsch, K. G. (2009). How do agents learn to behave normatively? Machine learning concepts for norm learning in the Emil project. In B. Edmonds & N. Gilbert (Eds.), *Proceedings of the 6th Conference of ESSA*.

Macy, M., & Sato, Y. (2002). Trust, cooperation, and market formation in the U.S. and Japan. *PNAS, 99*, 7214–7220.

Neumann, M. (2006). Emergence as an explanatory principle in artificial societies. Reflection on the bottom-up approach to social theory. In F. Squazzoni (Ed.), *Epistemological aspects of computer simulation in the social sciences* (Vol. LNAI 5466, pp. 69–88). Berlin: Springer.

Neumann, M. (2008). Homo socionicus: A case study of simulation models of norms. *Journal of Artificial Societies and Social Simulation, 11*(4), 6.

Neumann, M. (2009). Dissecting the BOID perspective on norms. In B. Edmonds & N. Gilbert (Eds.), *Proceedings of the 6th Conference of ESSA*.

Plotkin, H. C. (2003). We-intentionality: An essential element in understanding human culture? *Perspectives in Biology and Medicine, 46*(2), 283–296.

Ritzer, G. (2007). *Sociological theory* (7th ed.). Boston, MA: McGraw Hill.

Roth, A. S. (2004). Shared agency and contralateral commitments. *The Philosophical Review, 113*(3), 359–410.

Saam, N. J., & Harrer, A. (1999). Simulating norms, social inequality, and functional change in artificial societies. *Journal of Artificial Societies and Social Simulation, 2*(1), 2.

Schelling, T. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology, 1*, 143–186.

Schmid, H. B. (2008). Plural action. *Philosophy of the Social Sciences, 38*(1), 25–54.

Searle, J. (1983). *Intentionality: An essay in the philosophy of mind*. New York: Cambridge University Press.

Searle, J. R. (1995). *The construction of social reality*. London: Penguin.
Tomasello, M. (2009). *Why we cooperate*. Cambridge, MA: MIT Press.
Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*, 675–735.
Tomasello, M., & Rakoczy, H. (2003). What makes human cognition unique? from individual to shared collective intentionality. *Mind and Language, 18*(2), 121–147.
Tomasello, M., & Rakoczy, H. (2007). The ontogeny of social ontology. In S. L. Tsohatzidis (Ed.), *Intentional acts and institutional facts: essays on john Searle's social ontology*. Dordrecht: Springer.
Tuomela, R. (2007). *The philosophy of sociality: The shared point of view*. Oxford: Oxford University Press.
Tuomela, R., & Miller, K. (1988). We-intentions. *Philosophical Studies, 53*, 367–389.
Xenitidou, M., & Elsenbroich, C. (2010). Construct validity of agent-based simulation of normative behaviour. *The International Journal of Interdisciplinary Social Sciences, 5*(4), 67–80.
Zaibert, L. A. (2003). Collective intentions and collective intentionality. *American Journal of Economics and Sociology, 62*(1), 209–232.

Chapter 6
The Relational Foundation of Norm Enforcement

Christine Horne

What Are Norms?

Norms are rules, about which there is some degree of consensus, that are socially enforced. Norms therefore overlap with, but are distinct from, internal states such as values or morals. The key element of norms that distinguishes them from internal states is their social nature—the fact that they are enforced externally by individuals.

Further, norms are not simply descriptive. That is, norms are not equivalent to the sum of behaviors in a group—the frequency or the typicality of a behavior. Patterns of behavior may provide information about what the norms are. But, in order for norms to exist, sanctioning must occur.

This conceptualization of norms means that in order to fully understand them, we have to explain why they are enforced. It is not obvious why people punish. Sanctioning is costly. It can take time and effort. It can be embarrassing and provoke retaliation. So why do it? Existing explanations focus on the characteristics of behavior (see, for example, Coleman, 1990) and the human brain—finding evidence, for example, that harmful behaviors make us angry and motivate us to punish (Fehr & Gächter, 2002). These explanations contribute to our understanding of sanctioning but still leave unanswered questions. If all that mattered was the characteristics of behavior and the anger it provoked, then we would expect the same behavior to be treated the same way in every time and place. But, it is not. Just as the state enforces laws more in some neighborhoods than in others, so do people enforce norms more in some groups and contexts than others. Harmful behaviors are sometimes ignored, and harmless behaviors are sometimes punished. In some social environments, even cooperative behavior may be sanctioned

C. Horne (✉)
Department of Sociology, Washington State University, Pullman, WA, USA
e-mail: chorne@wsu.edu

M. Xenitidou and B. Edmonds (eds.), *The Complexity of Social Norms*, Computational Social Sciences, DOI 10.1007/978-3-319-05308-0_6,

(Herrmann, Thöni, & Gächter, 2008). Something more than simply the harm caused by a behavior or emotional reactions to it must be driving enforcement.

The view of norms as social (rather than as something that exists inside individuals' heads or as simply patterns of behavior) further implies that they are a group-level rather than an individual-level phenomenon. They emerge in groups and are maintained by groups. Thus to understand norms, we need to study not just sanctioning, but sanctioning within groups.

Evidence for the Relational Foundation of Norms

I argue that understanding of social relationships is essential for explaining norm enforcement. I have developed a set of theoretical predictions and tested them in a series of laboratory experiments. Below I describe the theory and the experimental evidence that shows how social relationships affect sanctioning and how sanctioning can affect social relationships.

Social Relationships and Norm Enforcement

I focus on one key characteristic of social relationships—dependence—as well as on two other social factors—metanorms and metanorm expectations. For each of these three factors I present the theoretical argument, briefly describe the experiments testing the theory, and summarize the results.

Dependence and Sanctioning Benefits

Dependence refers to the extent to which an individual values his or her relationship with another person and the goods that he or she can get from that relationship (Emerson, 1962, 1972; Molm, 1997; Molm & Cook, 1995).[1] The more that individuals in a group depend on one another, the more interdependent they are and the more cohesive the group (Emerson, 1962, 1972).

What is the connection between dependence and sanctioning? Researchers frequently assume that people will punish behaviors that cause harm (see, for example, Coleman, 1990). On this view, enforcing norms produces direct benefits (a reduction in antisocial behavior) for all those affected by the target behavior. Those direct benefits provide an incentive to sanction. But when group members are interdependent, enforcing norms can also produce indirect benefits (Horne, 2004).

[1] The definitions of dependence and cohesion used here are drawn from Emerson and Molm. Their work is part of a larger body of research on exchange developed by theorists Blau (1964), Homans (1974), and Kelley and Thibaut (1978).

This occurs when the gains that people experience as a result of deviance being discouraged increase their ability to exchange. In other words, people benefit directly when harmful behavior is punished. They also receive indirect benefits when they interact with others who have gained from the punishment of harmful behavior. This is because when individuals are dependent on those around them, their well-being is tied to the well-being of those others. They benefit when those with whom they interact have something to offer. If the other has few resources, the individual will not be able to gain much through exchange with that other—no matter how much he or she needs what the other has, the other will not have much to give. Thus individuals benefit when they are not personally victimized; they also gain when their neighborhood is safe and secure. Individuals prosper when they are not cheated; they also profit when levels of cheating are relatively low.

The fact that an individual's well-being is connected to the well-being of others means that the gains associated with sanctioning may be larger than they appear on their face (Horne, 2004). It also means that as patterns of social relationships change, the benefits associated with sanctioning shift, and sanctioning behaviors shift as well—even if the target behavior and the harm that it causes remain the same (Horne, 2008).

Further when individuals are dependent on others, they may enforce norms that benefit those others, even if they personally would prefer a different norm to be in effect. In some situations, everyone agrees on the harm caused by a behavior. In others, people have different interests in a behavior (or different understandings of its consequences). When this occurs, dependence relations can lead people to enforce norms they do not prefer.

In other words, even when there is a rule forbidding a harmful behavior and people disapprove of it, punishment of that behavior will vary. When social relationships are strong sanctioning will increase; when they are weak sanctioning will decline (Horne, 2001, 2008). We cannot assume that people will react negatively to harmful behavior or that a normative rule will be enforced consistently across social settings. Instead, the extent to which norms are enforced varies with the structure of social relationships (Horne, 2007, 2008). Accordingly, we would expect norms to grow and fade as social relationships shift.

To test these ideas I conducted two experiments using a norms game (Horne, 2008). Four subjects played a public goods game in which they had opportunities to contribute to one or more group funds. Individuals who contributed bore the costs, but all group members benefitted. Thus each individual hoped that others would contribute but also experienced the temptation to free-ride. Following each contribution decision points in the group fund were distributed to group members. Then all group members engaged in exchange—each person made decisions about how many of their points they wanted to keep, and how many they want to give to each other group member. They could adjust the number of points they gave to others based on whether those others had contributed to the group fund. The difference in what subjects gave to those who donated to the group compared with those who had the opportunity to do so but did not constituted a sanction. If participants gave more points to those who contributed than to those who did not, then they were enforcing a norm-favoring contribution. Participants played this game for a large number of rounds.

I manipulated the benefits associated with contributing to the group (and accordingly, the benefits group members would receive if sanctions encouraged people to contribute). When benefits were low, an individual's contribution produced only a few (two) points for each group member. When benefits were high, the individual's contribution produced a larger number of points (six). Participants gained more from others' contributions in the large than small benefit condition and therefore presumably had more interest in seeing that failures to contribute were sanctioned.

I also manipulated the extent to which group members were dependent on each other. I did this by varying the value of points that individuals received from others compared to the value of points in the individuals' own personal accounts. In the low-dependence condition, points that an individual received from others were worth the same as their own points. Participants could do just as well on their own as they did interacting with others. In the high-dependence condition, points that the individual received from others were worth three times their own points. Participants earned more points if they exchanged with others than if they did not.

The results showed that the size of the benefits associated with donating to the group fund (or the harm to group members when donations were not made) did not, in and of itself, affect sanctioning (Horne, 2008). That is, sanctions were not necessarily stronger when the consequences of the target behavior were larger. Rather, the consequences of the behavior interacted with the level of interdependence in the group such that sanctioning was greatest in groups in which the consequences of donating to the group were large *and* group members were highly interdependent.

Further, variation in the structure of dependence relations within a group affected patterns of sanctioning (Horne, 2008). In some conditions, subjects had conflicting interests in the group funds. In those conditions, individuals who were dependent on other group members tended to enforce norms that benefitted those others, rather than the norms they personally preferred. (Interestingly, although subjects in interdependent groups enforced norms preferred by others, they did not necessarily follow them.)

The results are consistent with the argument that interdependence among group members magnifies the benefits of sanctioning, in turn strengthening enforcement. They demonstrate that social relationships matter for norm enforcement.

Metanorms

Norm enforcement differs from punishment. Anybody can punish anyone for anything. But norm enforcement requires some element of consensus within the group. Consensus might arise if the target behavior affects all group members in the same way. If everyone has the same interest in a behavior, we would expect to see all those in the same situation react similarly. But consensus may have a more social component. Individuals care about what others think of them. They want others to cooperate with them. To encourage them to do so, the individual needs to demonstrate that he or she is a good person with whom to interact. Therefore, people will try to behave in ways that will maintain relationships and lead others to treat them positively rather than negatively.

What can people do to demonstrate that they are good exchange partners and good group members? One thing they can do is follow norms. When individuals obey group norms, they demonstrate that they know how to behave. Their actions provide evidence of their reliability and trustworthiness.

In addition to following norms, people can enforce them. If an individual punishes behavior that other group members would like to see punished, then he or she is demonstrating that he or she understands what the group norms are. Further, he or she is providing evidence of his or her commitment to the norm. He or she is establishing that he or she is willing to bear personal costs to enforce it (Posner, 2000). And he or she is showing that he or she is not just a poser—imitating others for the sake of popularity (Centola, Willer, & Macy, 2005). People demonstrate a commitment to honesty when they themselves are honest, but they also demonstrate that commitment when they punish deceit, blow the whistle on bad behavior in the workplace, and so forth.

Individuals who want to be treated well need to consider how their actions—including their sanctioning behavior—will be seen by those around them. When thinking about enforcing a norm, they will take into account the costs (potential retaliation, emotional discomfort, and so forth) and the benefits (including a reduction in deviant behavior). But they will also consider how others are likely to view their sanctioning activity. They anticipate potential reactions. In other words, they pay attention to metanorms (Horne, 2001).

Metanorms are a particular kind of norm that regulate sanctioning (Axelrod, 1985; Coleman, 1990). Like norms, they are socially enforced. The incentives provided by metanorms are selective—given only to the sanctioner. While the benefits of sanctioning are experienced by everyone, thus tempting people to free-ride, metanorms produce consequences only for the person who imposes the punishment. So, only the person who actually sanctions is rewarded.

Why do people enforce metanorms? Why do they reward sanctioners? Because no one, including sanctioners, wants just a fair-weather friend. Everyone maintains relationships that support them at some times but make demands on them at others. If people want to maintain relationships, then they stick with them through the profitable times as well as those times when the other has little to offer. The same is true of relationships with sanctioners. Individuals provide support to the sanctioner because they value the relationship. If they fail to be supportive, and some other group member remains loyal, then in the future the sanctioner is likely to defect to this more faithful acquaintance. The motivation to maintain ties is stronger when a relationship is valued. The more dependent people are on the sanctioner, the more they will want to support him or her. Thus dependence between group members increases the support given to sanctioners; it strengthens metanorms. In turn, metanorms affect norm enforcement (Horne, 2001, 2004).

To test these ideas, I conducted four experiments using a metanorms game (Horne, 2001, 2004, 2007; Horne & Cutlip, 2002). In this game, a computer-simulated thief stole from group members. Each time an individual was the victim of a theft, he or she could decide how to respond—whether to punish the thief or not. Group members also had opportunities to exchange with one another.

They could express their approval or disapproval of a victim's sanctioning decision by giving that person more or fewer points. The more points that participants gave to victims who punished the thief relative to those victims who did not punish, the stronger the metanorm-favoring punishment.

As in the norms game, I manipulated how dependent group members were on each other. I also manipulated the costs and benefits of sanctioning to see if metanorms could encourage people to sanction even when doing so imposed more costs than benefits on the group.

The results showed that when people were interdependent, they gave more support to sanctioners. They gave larger rewards to those who punished relative to those who did not. That is, interdependence strengthened metanorms (Horne, 2001, 2004, 2007; Horne & Cutlip, 2002).

In turn, when metanorms were strong, people were more likely to sanction. Groups with stronger metanorms had higher rates of punishment (Horne, 2001, 2004, 2007; Horne & Cutlip, 2002).

Further, metanorms encouraged people to sanction even when the costs of doing so were high and the benefits low (Horne, 2007). In fact, as the costs of sanctioning increased, the rewards given to sanctioners increased as well. Even when punishment was so costly that it produced an aggregate loss for the group, people who were highly dependent on each other rewarded such punishment. This encouragement in turn increased the rates of sanctioning in the group.

These results show that the same behavior that causes the same harm will be treated differently depending on the structure of social relationships. A behavior may be punished in one social environment and not in another. This is not because people are any less disapproving of the behavior. Rather, it is because the social relationships that support sanctioning are weak. Further, in groups in which members are dependent on each other, people may provide support to sanctioners that encourages them to punish deviance even when doing so is both individually and collectively irrational.

Metanorm Expectations

Rewards encourage sanctioning. But even anticipation of others' likely reactions may affect punishment decisions. Because people want to be rewarded, they try to determine what behaviors others would like to see punished and the punishment efforts that others will view positively. Thus, in addition to actual rewards and punishments, people's expectations about what sanctioning behaviors others are likely to approve also drive sanctioning (Willer, Kuwabara, & Macy, 2009).

The problem for the individual is that it is not always clear exactly what others want. In forming expectations therefore, people rely on a number of clues. One clue is the harm caused by a behavior. It is reasonable to think that if a behavior hurts others, those others would like to see it punished. Another clue is the frequency of behavior (Horne, 2009b). If the individual sees many others engaging in a particular

behavior, he or she might well conclude that others approve of the behavior and would disapprove of aberrations. People may also rely on the characteristics of the setting (Horne, 2010). It is widely known, for example, that informal control of criminal and deviant behavior varies across neighborhoods. Neighborhoods with certain characteristics (high poverty, low stability, and so forth) have lower levels of informal control and higher rates of crime. Why is this? One possibility is that people in those neighborhoods do not expect that others will support their sanctioning efforts.

I conducted several experiments that test the arguments that existing patterns of behavior and characteristics of the setting create expectations about others' potential reactions and that those expectations affect sanctioning decisions (Horne, 2009b, 2010).

The first two experiments test the argument that people use the typicality of behavior as a clue to help them anticipate others' potential reactions (Horne, 2009b). I created an expectations game in which each subject participated in a group with seven simulated actors. The actors took turns making a particular choice—the choice between X and W. This decision was as arbitrary as it sounds. Subjects literally had to choose between the two letters. The X–W choice had no consequences in and of itself. It had no association with status, aesthetic judgment, norms, or any other evaluation outside the lab. The point was to create an artificial behavior with no or as little as possible existing social meaning. The only factor that might make X or W more socially salient was the number of actors in the group who chose it.

The actors made their X–W choices one at a time. The subject went seventh. This meant that he saw all but one actor's choice before making his own. After the last actor made his X–W decision, everyone was able to react to each other's choices by giving them points. This time the subject went first. He had to make his sanctioning decision without knowing what anyone else would do. But, he knew that other people would be making their sanctioning decisions after him. And if others reacted negatively to him, he would have fewer points to take home at the end of the experiment.

The second experiment was the same as the first except that subjects were given information that made the X–W choice more socially meaningful. The experiment instructions said that research has revealed a surprising, yet consistent, finding—preferences for particular patterns of lines are associated with the number of friends people have. Those who prefer one category of line tend to have more friends; those who prefer the other tend to have fewer friends. The line patterns were the letters X and W. In other words, subjects had exactly the same choice to make as those participating in the first experiment. But this time they had information that their X–W choice might tell people whether they had lots of friends or only a few.

In the first experiment, the results show that behavior patterns had no effect on sanctioning. But, in the second experiment in which subjects had reason to think that their choices were socially meaningful, the typicality of behavior in the group did affect subjects' sanctioning decisions. Participants gave fewer points to those who made an atypical choice (Horne, 2009b).

I conducted a third study to examine the effects of setting on metanorm expectations and sanctioning (Horne, 2010). In this study, I showed college student participants pictures of a "good" neighborhood and a "bad" neighborhood. I told them to imagine that they were in the neighborhood and saw a crime being committed. I asked them how likely they would be to do something to try to stop the crime and, if they did so, how they thought others in the neighborhood would react.

The results showed that experimental subjects shown pictures of "bad" neighborhoods said that they were less likely to intervene to stop a crime than those shown pictures of "good" neighborhoods. These responses were completely explained by subjects' expectations regarding how positively or negatively other residents were likely to react to their intervention efforts (Horne, 2010). That is, characteristics of the setting (the neighborhood) affected metanorm expectations. Those expectations explained subjects' stated willingness to sanction. These results are consistent with the patterns of informal sanctioning across urban neighborhoods. Criminological research has long shown that people in "bad" neighborhoods exercise less informal control than those in "good" neighborhoods (see, for example, Bursik & Grasmick, 1993; Sampson & Groves, 1989; Sampson, Raudenbush, & Earls, 1997). The experimental results described here show that metanorms explain the link between neighborhood characteristics and sanctioning in the lab, suggesting that metanorm expectations may help to account for this link in neighborhoods.

Thus there is evidence that existing patterns of behavior and the characteristics of a setting may affect metanorm expectations and, in turn, sanctioning. There are other possible sources of clues as to the sanctioning actions that others will support. Individuals may also use others' sanctioning behaviors as sources of information. Consider, for example, nations' commitment to the International Criminal Court (ICC). The ICC enforces international human rights norms. When nations committed to the ICC, they made a commitment to the enforcement of those norms. Why did nations commit? One reason is that they considered the likely reactions of other countries on which they were dependent (Goodliffe, Hawkins, Horne, & Nielsen, 2012). The rhetoric of those other countries provided some information about likely reactions, but their actual commitment to the ICC provided even more. Thus a nation thinking about whether or not to commit to the ICC would consider whether other countries on which it was dependent had committed. As nations made commitments, the calculations of the uncommitted countries tied to those nations shifted. Very quickly nations' commitment decisions in conjunction with patterns of interdependence between countries produced increasing numbers of commitments (Goodliffe et al., 2012).

The studies described above provide evidence that metanorm expectations matter. If expectations of reactions affect sanctioning decisions, then expectations may become self-fulfilling prophecies. In particular, if group members' expectations are wrong—if they misperceive the behaviors others disapprove and would like to see sanctioned (see, for example, Perkins, Haines, & Rice, 2005)—then they will punish the wrong behaviors. But as they do so, they will create the norms they thought existed.

Norm Enforcement and Social Relationships

The series of studies described focuses on the effects of a structural feature (the characteristics of relationships within groups) on the emergence of norms (norm enforcement) within a group. But, norm enforcement also affects social relationships. That is, the causal arrow also goes the other way. As group members sanction, they hope that others will support their efforts. But they are taking a chance. They do not know for sure how others will react. As others actually reward sanctioning efforts, the relationships between the sanctioner and other group members become stronger. People place increasing value on their relationships (Horne, 2000).

Thus over time, as people enforce norms, relationships become stronger. This does not mean that the sanctioner's relationship with the deviant becomes stronger. Researchers often focus on the relationship between the sanctioner and the deviant—identifying features of that relationship that make sanctioning less likely and examining how sanctioning affects the relationship. Such work shows that strong relationships between deviants and potential sanctioners tend to dampen punishment. Here the focus is on the other group members. As group members anticipate support for punishing and as they provide support to others who sanction, they strengthen relationships with each other. Thus enforcing norms can make groups more cohesive, increasing the likelihood of future norm enforcement.

I conducted a study to test this dynamic (Horne, 2000). In particular, I examined how strengthening an alternative enforcement institution (the legal system) affected the informal controls enforced within groups. The legal system was operationalized as an agent that was supported by taxes collected from the group. The higher the taxes, the stronger the legal system. Further, the stronger the legal system, the lower the personal costs to any individual who turned to it to punish deviance. In contrast, enforcing norms personally was always directly costly to the individual. Individuals who personally punished deviance experienced costs. But when social relationships were strong, others helped to offset those costs through enforcing metanorms (rewarding those who punished). In contrast, when social relationships were weak, sanctioners received little support from others.

I found that as people used the legal system more—and bore lower personal costs for sanctioning—they also received less support from other group members (Horne, 2000). Over time, relationships weakened. People placed less value on their social relationships. In contrast, in conditions in which the legal system was weak, people engaged in more personally costly punishment and received more rewards from others. In turn, they placed more value on their relationships.

This finding is consistent with the work that shows that certain types of exchanges between actors can strengthen relationships (Lawler, 2001; Lawler, Thye, & Yoon, 2000). Norm and metanorm enforcement appear to involve interactions that similarly strengthen relationships. When people enforce norms, they hope that others will support them, but do not know for sure if they will. When others reward them for their efforts, they gain greater confidence in their relationships. Groups become more cohesive.

Summary

In sum, the structure (characteristics) of social relationships within a group affects sanctioning (Horne, 2009a). Sanctioning and support given to sanctioners in turn affect the characteristics of those relationships. That is, social structure affects individual enforcement efforts which in turn affect social structure. If this dynamic continues uninterrupted, we would expect to see groups become more and more controlling. Groups would be increasingly likely to enforce norms even if doing so provided few benefits. Norm enforcement would be strong. In contrast, if group members are not dependent on each other, if they do not value their relationships, then they are unlikely to enforce norms. It will be difficult for the group to achieve collective goals.

This dynamic of strong relationships facilitating norm enforcement which in turn strengthens relationships can be interrupted when outside institutions or events weaken people's dependence on each other. As in the study described above, increases in government involvement may weaken group members' dependence on each other for the punishment of deviant behavior. Many social institutions have the potential to weaken interdependence. When the law provides a substitute, cheaper source of control, individuals have less need of each other. The Internet weakens dependence on local social relationships for information. Employment law weakens the dependence of individuals on fellow union members. Employer-provided benefits weaken dependence on mutual benefit associations that in the past provided individuals and their families with security. Such social changes can affect the extent to which people are dependent on one another and, in turn, their sanctioning decisions.

Thus exogenous changes can have unexpected effects. If they provide a substitute for goods that people formerly worked together to provide, then they weaken people's dependence on each other. In turn, when people value their relationships less, they give less support to sanctioners. Metanorms are weaker. Norm enforcement declines. Norms lose their power.

Cumulating Theoretical Understanding of Norms

Dynamic approaches to studying norms allow for endogenous change; norms can evolve without external inputs. My work primarily focuses on how characteristics of social relationships (at the macro-level) affect the norm-related behaviors of group members (at the microlevel). I also have evidence regarding the effects of individual behaviors on characteristics of social relationships. Thus individuals are both affected by and affect the larger environment (see Andrighetto et al. and Burgemeestre et al., this volume, for alternative approaches to thinking about macro–micro-links). Even if normative rules remain constant, enforcement of those rules can change. Patterns of enforcement shift. Norms grow and fade with enforcement of the rule. Some norm change occurs endogenously; some change can be triggered by exogenous factors.

Other dynamic approaches similarly focus on the interplay between characteristics of the situation and individuals. Ostrom, Gardner, and Walker (2006), for example, identify structural features common to groups that have successfully solved collective problems (see also Janssen and Ostrom this volume, focusing on information as a key structural feature of groups). Bicchieri (2006) focuses on features of the environment that make norms salient—how the environment affects individuals' expectations regarding what others are likely to do and what those others expect, in turn affecting what the individual does. Rather than see individual internal states as immutable, Bicchieri sees them as shifting in response to the social environment. Andrighetto et al. (this volume) similarly discuss the interplay of social factors and individual internal states.

The "social norms approach" also emphasizes the effects of the larger environment on individuals. For example, research shows that patterns of behavior (such as drinking in college campuses) can affect students' perceptions of how much their fellow students drink and how favorably those students see drinking. These perceptions affect the individual student's own drinking behavior, which in turn contributes to perceptions. This pattern can be interrupted by providing students with accurate information about what their fellow students actually think about drinking.

Other dynamic approaches to norms focus primarily on the intersections of individual behaviors. These approaches embed behavioral assumptions in agents who then act. Individual behaviors affect others' decisions. Thus individuals are affected by the behavioral rules they are programmed to follow and the behaviors of those around them. Individual actions intersect to produce macro-level patterns of behavior (for an example of this approach see Elsenbroich, this volume). Interactions may lead to a variety of macro-level outcomes—equilibrium, continual change, cycling between different macro-level patterns, and so forth.

Thus some dynamic approaches focus on macro-level features of the environment, their effect on individuals, and the effect of individual behaviors on those macro-level features. Others focus on how the interplay of behaviors by actors following simple behavior rules produces macro-level patterns. The difference between these two approaches is that the first embeds influences on individuals in the social environment, while the second produces individual behavior by embedding assumptions in actors. This distinction is not as great as it may appear. At a conceptual level, structural constraints can be converted into internal states of agents or vice versa. For example, an assumption that individuals have a taste for conformity might produce the same kinds of behaviors as a social environment that restricts choices. An assumption that individuals have a taste for uniqueness might produce behaviors similar to those we would see in a social environment that encourages innovation. Though subtle, this distinction may nonetheless affect how researchers think about norms. At a practical level, because agent-based models highlight characteristics of actors and the distribution of actors of different types, the most obvious type of intervention is to change the characteristics of actors or their distribution. But because outcomes are the consequence of many interactions, it is difficult to predict what the outcome of a particular change would be. Further, it is easier to

change our assumptions about actors than it is to change actors themselves. In contrast, a structural approach highlights structural interventions as a way to change norms. Whatever the characteristics of individual actors, certain kinds of structures foster norm enforcement more than others.

While we have learned much about norms, there are still many unanswered questions. What might help us as we continue to study the emergence, change, and decline of norms?

Values and Expectations

Norms may be effective because they are internalized into the individual's value system. They may also be effective because they shape individuals' perceptions and expectations that in turn affect their behavior. Researchers differ in the extent to which they emphasize these two mechanisms (for a related discussion see Xenitidou & Elsenbroich, 2010). To some extent, agent-based approaches locate norms in the individual, while structural/situational approaches locate them in the larger environment.

But research is often not clear about the extent to which norms produce individual values or strategies for action (Yamagishi, Hashimoto, & Schug, 2008). For example, research suggests that cultural variation is reflected in individual internal states (see, for example, Haidt & Graham, 2007; Haidt, Koller, & Dias, 1993; Markus & Kitayama, 1991). Much works seem to assume that norms have been internalized so that individuals in different cultures adhere to different values and therefore behave differently. It is possible that individuals carry cultural tastes and preferences into the lab with them. But it is also possible that they bring expectations about others into the lab. Thus their behavior may reflect individual values or it may reflect strategies based on understanding of a society's norms (Bicchieri, 2006; Yamagishi et al., 2008).

The fact that research does not always explicitly distinguish between the two possibilities is a problem because internalized values and expectations about the social world may be the result of different causal factors and mechanisms and may have their effects through different mechanisms. Values are thought to be relatively stable and carried in the individual from one context to another. Expectations are more likely to be formed in situations and to change as the individual moves across social contexts. Researchers need to be clear about these two possibilities in order to collect data evaluating their contribution. Research questions suggested by a focus on internal states will likely be different from those suggested by a focus on the external environment. In the context of neighborhood crime, for example, criminologists have tried to explain why crime rates are higher in neighborhoods with some characteristics than others. One explanation is that norms differ. But, when researchers talk to people they do not find evidence that people in poor neighborhoods have different values than those in rich neighborhoods (Kornhauser, 1978: 214–221). As described above, however, it may well be the case that people in "bad"

neighborhoods have different expectations of their neighbors than people in "good" neighborhoods. People in both kinds of neighborhoods may disapprove of crime (have the same values), but they may have different expectations about the extent to which others disapprove of crime and will try to do something about it—and therefore different norm enforcement patterns. By distinguishing between these two possibilities, we might be better equipped to understand the relation between norms and crime across communities.

Behavior is likely due to some combination of individuals' preferences (what the individual wants, cares about, and so forth) and individuals' perceptions of others' preferences (what others want and will approve of). Similarly, norm enforcement is likely affected by individuals' views of behaviors and their perceptions of what others would like to see punished. Thus both values and expectations are likely to matter. Norms may be internalized so that they become values as well as norms. But blurring the distinction between the two makes cumulating theoretical knowledge about norms difficult.

In addition to identifying individuals' expectations of others, we also need to do more work to understand how those expectations develop and change. While researchers have begun to look at this issue (see, for example, Bicchieri, 2006), there is still much that we do not understand.

Substantive and Abstract Norms

Researchers study norms at different levels of abstraction. Conceptualizing norms in terms of games such as the ultimatum game or the social dilemma game has produced research that has contributed greatly to our understanding of norm enforcement. There are many benefits to focusing on basic theory and abstract norms, including that doing so contributes to the cumulation of knowledge. But such approaches may limit the range of norms that we consider. It may be useful, therefore, to also study substantive norms.

For example, consider American norms governing race relations. The explicit norm is that we should be color-blind, that race does not matter. Despite this explicit norm, many Americans think that other Americans disapprove of intimate relations across racial lines (Dave Thomas Foundation for Adoption, 2002). These perceptions are inconsistent with both what people say they support and what they think the explicit norm is. So in the United States explicit norms have shifted from making distinctions between racial groups to saying that race should not matter. At the same time, there are people who are aware of the explicit norm of color-blindness, whose own views may or may not be consistent with that norm, and who believe that others do not support the explicit norm. Although the explicit norm is that race should not matter, segregation persists. Such complex norms raise questions that might not be raised if our research is limited to standard games. In this case it suggests that norms may be more or less explicit and that explicit and implicit norms may differ and may vary in how they change over time.

Lab and Field

I tested my theoretical ideas primarily using lab experiments. Experiments are very useful for testing theories because they provide strong evidence of causation. Many norm scholars rely on experiments and in particular on standard games. Others use data from the field. For example, they use computer simulations to see if they can recreate real-world conditions in a simulation outcome. To the extent that simulation outcomes are consistent with real-world patterns, there is support for the theory embedded in the simulation. Similarly, to the extent that data from the lab and the field are consistent, we can have greater confidence in our theories (see, for example, Ostrom et al., 2006).

Lab experiments can be used to explicitly tests ideas suggested by results obtained in the field. For example, researchers have found that people in different societies play standard economic games differently. Further, they have found correlations between characteristics of the society and how people in those societies play standard games. In many ways these findings reflect those of survey researchers showing that values/attitudes vary across cultures. They show that market integration increases individuals' cooperative behavior in the ultimatum game (Henrich et al., 2001). They also show that variation in antisocial punishment across culture is correlated with the rule of law and religion (Herrmann et al., 2008). But the evidence is largely correlational. Further, the reasons for these correlations are not well understood. Standardized games are useful because they provide behavioral measures and because they facilitate making comparisons across studies and cultures. They suggest important insights into cultural and structural factors that may affect norms. Researchers could build on these findings to develop theory about just how structure/culture affect norms. To test such causal theories, one could manipulate cultures and structures in the lab and observe the norms that emerge. That is, it may be useful to depart from standard games and design experiments that will test the effects of societal level factors on norms.

Lab experiments also have their limits. For example, incentive structures in the lab are usually clear. In the field they are not. Yet norms emerge amidst uncertainty, ambiguity, and conflict. To fully take advantage of field settings, however, we need to develop ways to measure norms. While measuring norms in the lab is relatively straightforward, in the field it is more challenging. It is difficult to get accurate indicators of norms that people do not want to talk about. Researchers need to be able to measure explicit norms as well as the norms that people think exist but will not admit to believing themselves. We must also be careful to develop measures that distinguish between values and norms.

Conclusion

Researchers have learned much about the emergence of norms over the last 20 years. Yet, unanswered questions remain about norm enforcement and even more about norm content and norm change. To move our understanding forward we should take

advantage of the strengths of different approaches. In order to do so, researchers need to be clear about their assumptions so that those different approaches can build on each other's knowledge. In particular, we need to be clear about the distinction between evaluations that are social and those that are internal to individuals. We need good measures of theoretical concepts that can be used across settings. Finally, it will be useful to bring multiple methods to bear on the same theoretical problems. Computer simulations will allow us to examine complex interaction processes and their outcomes. Lab experiments will allow us to empirically test causal relations and mechanisms. And applications in the field will allow us to explore the applicability of our theories across settings. We will learn more through using multiple methods than any single approach alone. Taking advantage of the strengths of multiple approaches will contribute to the development of cumulative theoretical knowledge.

References

Axelrod, R. (1985). An evolutionary approach to norms. *American Political Science Review, 80*(4), 1095–1111

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Blau, P. M. (1964). *Exchange and power in social life*. New York: Wiley.

Bursik, R. J., Jr., & Grasmick, H. G. (1993). *Neighborhoods and crime*. New York: Lexington Books.

Centola, D., Willer, R., & Macy, M. (2005). The emperor's dilemma: A computational model of self-enforcing norms. *American Journal of Sociology, 110*, 1009–1040.

Coleman, J. (1990). *Foundations of social theory*. Cambridge, MA: Harvard University Press.

Dave Thomas Foundation for Adoption. (2002). National adoption attitudes survey. http://www.adoptioninstitute.org/survey/survey_intro.html. Retrieved on November 16 2010.

Emerson, R. M. (1962). Power-dependence relations. *American Sociological Review, 27*(1), 31–41.

Emerson, R. M. (1972). Exchange theory, part II: Exchange relations and networks. In J. Berger, M. Zelditch Jr., & B. Anderson (Eds.), *Sociological theories in progress* (Vol. 2, pp. 58–87). Boston, MA: Houghton Mifflin.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137–140.

Goodliffe, J., Hawkins, D., Horne, C., & Nielsen, D., (2012). Dependence networks and the International Criminal Court. International Studies Quarterly (Forthcoming).

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98–116.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65*(4), 613–628.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al. (2001). Cooperation, reciprocity, and punishment in fifteen small-scale societies. *AEA Papers and Proceedings, 91*(2), 73–78.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*, 1362–1367.

Homans, G. C. (1974). *Social behavior: Its elementary forms*. New York, NY: Harcourt Brace and World.

Horne, C. (2000). Community and the state: The relationship between normative and legal controls. *European Sociological Review, 16*(3), 225–243.

Horne, C. (2001). The enforcement of norms: Group cohesion and meta-norms. *Social Psychology Quarterly, 64*(3), 253–266.
Horne, C. (2004). Collective benefits, exchange interests, and norm enforcement. *Social Forces, 82*(3), 1037–1062.
Horne, C. (2007). Explaining norm enforcement. *Rationality and Society, 19*(2), 139–170.
Horne, C. (2008). Norm enforcement in heterogeneous groups: Sanctioning by majorities and isolated minorities. *Rationality and Society, 29*(2), 147–172.
Horne, C. (2009a). *The rewards of punishment: A relational theory of norm enforcement*. Stanford, CA: Stanford University Press.
Horne, C. (2009b). Metanorm expectations: Determining what to sanction. *Advances in Group Processes, 26*, 199–223.
Horne, C. (2010). Unpublished data.
Horne, C., & Cutlip, A. (2002). Sanctioning costs and norm enforcement. *Rationality and Society, 14*(3), 285–307.
Kelley, H. H., & Thibaut, J. W. (1978). *Interpersonal relations: A theory of interdependence*. New York: Wiley.
Kornhauser, R. R. (1978). *Social sources of delinquency*. Chicago: University of Chicago Press.
Lawler, E. J. (2001). An affect theory of social exchange. *American Journal of Sociology, 107*, 321–352.
Lawler, E. J., Thye, S., & Yoon, J. (2000). Emotion and group cohesion in productive exchange. *American Journal of Sociology, 106*, 616–657.
Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*, 224–253.
Molm, L. D. (1997). *Coercive power in social exchange*. Cambridge: Cambridge University Press.
Molm, L. D., & Cook, K. S. (1995). Social exchange and exchange networks. In K. S. Cook, G. A. Fine, & J. S. House (Eds.), *Sociological perspectives on social psychology* (pp. 209–235). Boston, MA: Allyn and Bacon.
Ostrom, E., Gardner, R., & Walker, J. (2006). *Rules, games, and common-pool resources*. Ann Arbor, MI: University of Michigan Press.
Perkins, H. W., Haines, M. P., & Rice, R. (2005). Misperceiving the college drinking norm and related problems: A nationwide study of exposure to prevention information, perceived norms, and student alcohol misuse. *Journal of Studies on Alcohol, 66*(4), 470–478.
Posner, E. (2000). *Law and social norms*. Cambridge, MA: Harvard University Press.
Sampson, R. J., & Groves, W. B. (1989). Community structure and crime: Testing social disorganization theory. *American Journal of Sociology, 94*(4), 774–802.
Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science, 277*(5328), 918–924.
Willer, R., Kuwabara, K., & Macy, M. (2009). The false enforcement of unpopular norms. *American Journal of Sociology, 115*, 451–490.
Xenitidou, M., & Elsenbroich, C. (2010). Construct validity and theoretical embeddedness of agent-based models of normative behavior. *International Journal of Interdisciplinary Social Sciences, 5*(4), 67–80.
Yamagishi, T., Hashimoto, H., & Schug, J. (2008). Preferences versus strategies as explanations for culture-specific behavior. *Psychological Science, 19*(6), 579–584.

Part II
Methods and Epistemological Implications of Social Norm Complexity

Chapter 7
Norm Emergence in Regulatory Compliance

Brigitte Burgemeestre, Joris Hulstijn, and Yao-Hua Tan

Introduction

In their daily activities, people and businesses are subject to all kinds of governmental regulations. There is a great variety in the setup of regulative systems that aim to enforce a subjects' compliance with these regulations (Burgemeestre, Hulstijn, & Tan, 2009b). A trend in regulatory compliance is to formulate legislation as the so-called open norms: norms which leave room for contextual interpretation about their implementation (Gribnau, 2008). Open norms are considered to be flexible and adaptable to changing and varying circumstances, whereas rules require constant amendment to meet changing circumstances (Ford, 2008). Instances of open norms are principles or regulatory objectives. Unlike rules, open norms do not describe in detail what is permissible or prohibited; they indicate what behavior is expected by abstract principles or general objectives. For example, a recipe may prescribe to add salt "according to taste." How much salt must be added depends on the cook and his guests. Through experience a cook learns what reasonable quantities of salt are appropriate for different kinds of dishes. So in general, open norms need to be interpreted and operationalized for a specific context. Interestingly, this interpretation and operationalization process itself is a normative process.

In this work, we discuss the emergence of norms in human practice. In particular, we discuss the emergence of normative interpretations of open norms in a specific domain. Emergence can be defined as the generation of macro-social properties by

B. Burgemeestre (✉)
Faculty of Economics and Business Administration, Vrije Universiteit, Amsterdam, The Netherlands
e-mail: c.b.burgemeestre@vu.nl

J. Hulstijn • Y.-H. Tan
Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands

M. Xenitidou and B. Edmonds (eds.), *The Complexity of Social Norms*, Computational Social Sciences, DOI 10.1007/978-3-319-05308-0_7,

micro-social behavior (Conte, Andrighetto, Campenni, & Paolucci, 2007). By interacting, individual agents may generate and establish norms and continue to spread them, until at some point they affect the macro behavior of society. We consider two types of agents: agents that implement norms into information systems (subjects) and agents that prescribe and audit norm compliance (regulators). Open norms assume that also subjects are able to interpret and implement legislation appropriately (Burgemeestre, Hulstijn, & Tan, 2009a, 2009b; Burgemeestre et al., 2009b). Whereas rule-based norms are quite straightforward to implement and require little interpretation, open norms need to be tailored to an agent's specific situation. Open norms may gradually become more rule-like by the addition of best practices and requirements (Ford, 2008). Furthermore, open norms need to be translated into concrete constraints or rules before they can be implemented in IT systems (Sadiq, Governatori, & Namiri, 2007). From the regulators' perspective open norms also require a different approach to norm enforcement. Compliance with rules can be simply checked; compliance with principles or general objectives requires interpretation of norms and an evaluation of the implementation in a certain context. The regulator thus needs to get an insight into how and why a company has implemented principles or general objectives as certain rules in its business processes to determine compliance (Burgemeestre et al., 2009b). Even though the norms are adapted to the unique circumstances of an agent, norm compliance has to be enforced in a fair and consistent way. Regulators can therefore also learn from best practices developed in the field and use them as benchmarks to evaluate future implementations of open norms.

When legislation is relatively new or applicable to a (rapidly) changing domain, both types of agents may lack experience and knowledge to work with the norms. Subjects need examples of appropriate norm implementations from regulators. Regulators on the other hand need input from the field to determine fair evaluation criteria to assess norm implementation. Regulators are thus awaiting experiences from the field, and the subjects are awaiting implementation guidance and criteria. To overcome this potential deadlock situation, we need to understand the emergence of norms to be able to accelerate both of these learning processes. Understanding these processes is relevant for research in social simulation (Savarimuthu & Cranefield, 2009) and also for the specification of normative multi-agent systems (MAS) with certain desirable properties. Through the study of examples of norm emergence in human practice, we intend to extend theories on norm emergence in MAS research. In this work we will therefore focus on normative theories from MAS research.

Current research in MAS addresses norm evolution and adoption either as norm emergence (bottom-up) or norm prescription (top-down) (Savarimuthu & Cranefield, 2009). We argue that an integrated approach is needed to study and model norm emergence more realistically: bottom-up processes are likely to be bounded by law(s), and top-down processes may be influenced by bottom-up forces. Especially for compliance with open norms, where implementations are instantiations of prescribed open norms (top-down), but where consensus emerges about how little or how much compliance effort is considered acceptable (bottom-up), it is important to address both processes simultaneously.

In this chapter we report on a case study of norm emergence in the implementation of income taxation in the Netherlands. Specifically we look at a regulation concerning "valuation of non-monetary income: private use of company cars." This act prescribes that employees using a company car should keep a sound and auditable kilometer administration when they want to apply for a tax reduction. Based on interviews and document research we construct two scenarios about norm emergence: (1) caused by introduction of new technology and (2) caused by context-specific problems with the implementation. Using literature from simulation studies conducted in the MAS field, we review the (normative) changes that have occurred and evaluate whether the current theories on norm emergence are capable of capturing the findings from the case study.

The remainder of the chapter is organized as follows: First, we discuss a selection of literature on norm emergence. Next, we lay out our research approach in the following section. Then, in the core of the paper, we describe processes of norm emergence encountered in the case study. Finally, we discuss whether the literature explains the findings of our case study and identify gaps.

Theories of Norm Emergence

For the implementation and enforcement of open norms, we observe that norms emerge through discussions about compliance between and among companies and regulators. The emergent process that we encountered exhibits top-down as well as bottom-up characteristics. Open norms are prescribed in a top-down fashion from regulator to company. Implementation on the other hand occurs in a bottom-up way, as companies propose their specific implementation to the regulator.

Literature with a top-down norm prescription view can be found in research on legal sources, work on compliance, and also work on electronic institutions (e.g., Aldewereld et al., 2006; Dignum, 2002; Vazquez-Salceda, Aldewereld, & Dignum, 2005). A relevant aspect of norm prescription that is little discussed in MAS research is the specification of open norms. Dignum (2002) observes that norms in regulations are (on purpose) specified at a high abstraction level to account for many different situations which may occur over time. Regulations are specified on a higher abstraction level than the level on which the processes and structure of the institution are specified. Norms thus need to be translated and adapted to a certain domain before they can be implemented. The chapter discusses different levels of abstraction. Institutions are considered to have a predefined objective and a set of values that direct towards fulfilling that objective. Attached to each value is a list of norms that contribute to that value.

> Norms contribute to a value if fulfilling the norm always leads to states in which the value is more fully accomplished than the states where the norm is not fulfilled (Dignum, 2002).

The set of norms as a whole "defines" the meaning of the value in the context of the institution. The behavior of agents who operate in the institution must be in line with the objectives, values, and norms defined by the institution. In order to check norm compliance and act on possible violations, the abstract norms have to be translated into concrete specifications of behavior. "Concrete norms pertain to actions that are described in terms of the ontology of the institution and from which therefore the meaning and effect is known or they pertain to situations that can be checked directly by the institute" (Dignum, 2002). Based on a categorization of abstractness Dignum indicates how translations can be made between the types of abstract norms and more concrete norms. For the translation from concrete norms to the implementation level, Dignum suggests that for each concrete norm a rule is needed that specifies either part of the norm enforcement procedures of the institution or triggers that signal a violation of the norm and the expected reaction of the institution.

Aldewereld et al. (2006) describe the implementation of a norm enforcement mechanism for electronic institutions that is based on detecting violations of norms and reacting to these violations instead of restricting agent behavior up forehand. The chapter proposes to use integrity constraints and dialogical constraints to implement such a mechanism. They point out that the implementation of norms requires an operational semantics. The declarative nature of norms is necessary for reasoning about what is considered legal or illegal, while the operational semantics defines how norms are to be implemented, i.e., what to do when norms are in fact violated. Norms that are described in abstract terms first need to be "contextualized" before they can be implemented. This contextualization consists of two steps: (1) interpret the abstract concepts of the norm and link them to concrete concepts used in the institution, using counts-as operators, and (2) adding procedural information and artefacts to the institution to allow enforcement of the norm.

Vazquez-Salceda et al. (2005) discuss the operationalization of institutional norms in MAS in more detail. Norms are categorized depending on the actors involved, verifiability of states and actions in norm expressions, and temporal aspects. For each aspect guidelines on the implementation of norms are proposed.

These implementation frameworks assume that prescribed norms lead to the adoption of certain goals, or control objectives, which are subsequently turned into constraints on actions or control measures. But in the case of open norms, ideas on norm implementation can originate bottom-up from the subjects who like to comply. For example, for security regulations, we observe that companies copy security measures from successful competitors. Works about social networks or emergence of social conventions, like Savarimuthu and Cranefield (2009), study the processes of emergence of social norms or conventions.

Savarimuthu and Cranefield (2009) provide an overview of recent work on distributed normative behavior in simulation studies. The distributed approach focuses on the bottom-up emergence and spreading of norms in a system. For social norms they propose a norm-life cycle model that consists of four phases: (1) norm creation, (2) norm spreading, (3) norm enforcement, and (4) norm emergence.

For each of the phases of the norm-life cycle model, they present an overview of the mechanisms which are found in norm simulation research. The mechanisms fall into the following categories: social power (punishment, leadership), learning (machine learning, imitation), reputation, off-line design, cognitive, emotion-based, dynamic and static social network typologies, and cultural and evolutionary aspects.

Savarimuthu and Cranefield (2009) consider two variants of norm emergence: (1) the spreading of a norm in an agent society, until the norm is recognized and followed by most agents, and (2) bottom-up creation of norms by cognitive agents. Such agents derive a "proposed norm" based on their cognitive ability or come up with an alternative norm (creation phase) and then help in the emergence of that norm (emergence phase).

Unlike such frameworks, which allow basically any behavior to emerge as a norm, in our example the emergent norms are constrained by a related abstract norm specified by an institution. Not only actions of individual agents influence the evolution of norms in a society, but also norms in a society affect individual agents. Conte et al. (2007) describe these processes and the micro–macro link in more detail. They define emergence as the generation of macro-social properties by micro-social behavior. Individual agents intentionally or unintentionally produce effects by creating the conditions for them to arise. Conte et al. also describe the downward causation in which an emergent, macro-social property generates effects at the lower level agents. Downward causation occurs in two ways: by a simple and a complex feedback loop.

In the simple loop the emergent effect retroacts on the lower level by determining a new property of the generating system. In the complex loop, the emergent effect determines new properties by means of which the effect is reproduced again, which includes two sub-processes: immergence and second-order emergence.

> Immergence is the process by means of which the emergent effect modifies the way of functioning of the generating system, affecting its generating rules or mechanisms in such a way that it is likelier to be reproduced. Second order emergence (or incorporation) is the process by means of which an emergent effect is recognized by the producing systems and by this means, the effect is likelier to be reproduced (Conte et al., 2007).

In this case of immergence agents accept the norm and implement it into their processes. When agents become aware of the effects, that they contribute to generate through this implemented norm, we speak of second-order emergence. Conte and Castelfranchi (1999) try to establish connections between social conventions and prescriptions and model norms as cognitive objects. Their hypothesis is that the emergence of norms is intertwined with the emergence of normative beliefs. Unlike the conventionalists who see norm emergence as a process of agents imitating observable behavior of other agents, they see emergence as a noncontinuous phenomenon. A social norm is seen to imply a belief that a given behavior is generally prescribed within a (agent) community. This behavior is executed because and as long as it is believed to be obliged. The consequence of conforming to a given conduct—believed to be prescribed—promotes the act of prescribing it, thus contributing to its spreading, i.e., reproducing it.

Research Approach

The literature review shows that there are MAS studies on social mechanisms that help to explain how norm emergence occurs, technical papers that describe mechanisms for the specification of norms within electronic institutions, and papers that make an attempt to describe the link between the micro (individual agents) and macro level (agent societies). What has, to our knowledge, not been described is restricted or bounded norm emergence, where individual agents are allowed to create and spread norms within the normative boundaries set by an agent institution or society. Whereas our notion of "bounded norm emergence" is based on observations of human practice, we think that such a mechanism can also be of interest for MAS research. Consider for example an online community in which individual users are allowed to collectively define their own norms as long as they do not conflict with the core values of that community (e.g., Wikipedia) (Goldspink, 2009). Insights into human practice might be used to develop better balanced control mechanisms to facilitate norm emergence and to define circumstances under which bounded norm emergence can occur.

In the next section we describe our case study on norm emergence in income taxation in the Netherlands. We do realize that the literature on norm emergence often focuses on the emergence of social conventions and that income taxation in itself is not a typical example of a social convention. However, we think that the processes associated with the spread and emergence of social conventions resemble quite closely the normative processes that we observed in the Dutch taxation approach. Therefore, we think that it is appropriate to apply theories of norm emergence in this setting. We consider the case study method a suitable approach to study norm emergence, because norm emergence is a multifaceted phenomenon that needs to be studied in its real-life context (Eisenhardt, 1989; Yin, 2003). In addition, the emergence of norms is a complex social phenomenon that is highly dependent on contextual variables.

The Dutch approach to (income) taxation is a representative example of bounded norm emergence. First the tax legislation is set up in such a way that it does not prescribe what to do for regulators and subjects in detail and therefore allows room for individual implementations and judicial decisions (Gribnau, 2008). Furthermore, the Dutch tax administration is experimenting with forms of responsive regulation (Ayres & Braithwaite, 1992; Braithwaite, 2007). Unlike traditional forms of enforcement, the responsive approach of Dutch tax is based on trust and cooperation and requires an active participation of companies. This approach to tax control resembles quite closely the characteristics of bounded norm emergence where individual agents (companies) are allowed some freedom in the implementation of abstract norms prescribed by an institution (Dutch Tax Administration).

In this chapter we intend to explore and describe in more detail how and under which circumstances (bounded) norm emergence occurs. We therefore zoom into specific scenarios of norm emergence in income taxation. Through a combination of interviews and document research, we gather data to construct two scenarios that each describe the relevant agents, changes to norms, and circumstances under which

the emergence occurred. In the discussion (at the end) we relate the results of the case study back to the constructs provided by the theories described above. We evaluate which aspects are covered by the literature and where extensions to theory are needed. The evidence that is gathered and the conclusions drawn in this initial research may be used to develop future norm emergence mechanisms and models.

Case Study: Norm Emergence Concerning Kilometer Registration

We conducted a case study with the Dutch Tax Administration in which we studied the phenomena of norm emergence. A case study is a form of qualitative empirical research, especially useful for generating hypotheses (Eisenhardt, 1989). The purpose of this case study is to consider the working hypothesis that in case of open norms (principle-based regulation or regulatory objectives), the interpretation and operationalization of norms for specific circumstances is itself a normative process, which can be understood as a form of norm emergence.

Here we study the adoption, implementation, and enforcement of a piece of legislation in practice, namely, the Inkomstenbelasting, in particular Article 13bis (LB, 1994): "Valuation of non-monetary income: private use of company cars." When an employee drives more than 500 km for private purposes, a percentage of 25 % of the value of the lease car is added to the employee's income, which results in higher income taxes. When employees can prove that they drive less than 500 km on a yearly basis for private purposes, the value of the company car is not seen as income. Employees do not automatically benefit from the regulation; they have to indicate explicitly that they apply for the tax reduction. They do that by filling out a form during a lease contract or when a new lease contract is signed. The Dutch Tax Administration communicates clearly and frequently on the regulation, and the regulation is well known under lease car drivers.

To prove that one drives less than 500 km for private purposes, employees are obliged to keep a kilometer registration. Article 21. C of the Uitvoeringsregeling Loonbelasting (executive regulation) prescribes that:

> The the kilometer registration should contain the following elements: brand and type of car, license plate number, period of ownership and information on the driving trips.

Currently, the law supposes the use of a qualified odometer, a measurement device for recording distances, which is already built into every car. For each trip (one-way, between two addresses) one should keep a record of the:

Date
Start and end position of the odometer
Departure and arrival address
Route driven, when it differs from the usual route
Character of the trip (private or business)
Number of private kilometers, driven during a business trip

A kilometer registration is considered to be sound when the recorded kilometers correspond with the claims made in the fiscal report. Therefore, to ensure completeness, the total number of kilometers driven (business + private) should correspond with the distance recorded by the odometer. In addition, to ensure validity, the reported business kilometers should correspond with (supporting) evidence recorded in secondary registrations, such as calendars, planning systems, garage bills, or even traffic fines when obtained during business hours. The requirement of auditability (Article 52-6 of the General Taxes Act of the Netherlands) of the kilometer registration states that the tax inspector should be able to audit and verify the claims made in the fiscal report. One should therefore take measures to ensure that the source data and refinements and aggregations on that data are securely recorded and can be verified by an audit trail. For example, the details of each trip should be recorded, and not only the total amount of private kilometers. For people that drive to varying locations, registering all details of all trips can become quite an administrative burden. Some people may therefore give an estimation of their driven kilometers instead of an exact report of their trips. This is not considered as valid evidence and counts as a violation of the law. However, as fraud can be financially very attractive, incompliance can also be caused by deliberately under-reporting the number of kilometers. In the case of an incorrect kilometer registration, the lease car driver receives a fine and the income taxation is recalculated.

In this chapter we focus on the emergence of norms that occurs when specific groups of agents try to implement the abstract legislation in practice. Although at first sight the legislation seems detailed and allows little room for different interpretations, we do encounter quite some variations of kilometer registration implementations in practice. For instance, some people keep an old-fashioned logbook in a paper notebook; other people try to use their GPS smartphones or car navigation software to assist them with a computer-generated kilometer registration; and some people propose to use specific kilometer registration modules built into company planning software. After all, they argue, when you know all visits to business destinations from the planning module, that the number of private kilometers is also known by distracting this amount from the total amount of kilometers. Other variations concern the responsibility. In some cases registrations are kept by the company; in others it is the sole responsibility of the car owner. These initiatives are developed by companies and citizens. There is yet no guarantee that the tax office will consider these innovative registrations as valid, in a legal sense. Usually, individual variations in the implementation of a regulation do not bring about processes of norm emergence that affect the legislation. However there can be events that require adaptation of the legislation. In the kilometer administration case we see that under political pressure to reduce administrative burden, the Dutch Tax Administration, technology providers, software providers, branch organizations, and companies are together creating norms in order to establish technical standards and guidelines for some of these more innovative registrations to be considered legal and valid.

The process is shown in Fig. 7.1. Norms emerge when companies try to implement the legislation in practice and consider the use of new technologies or copy the

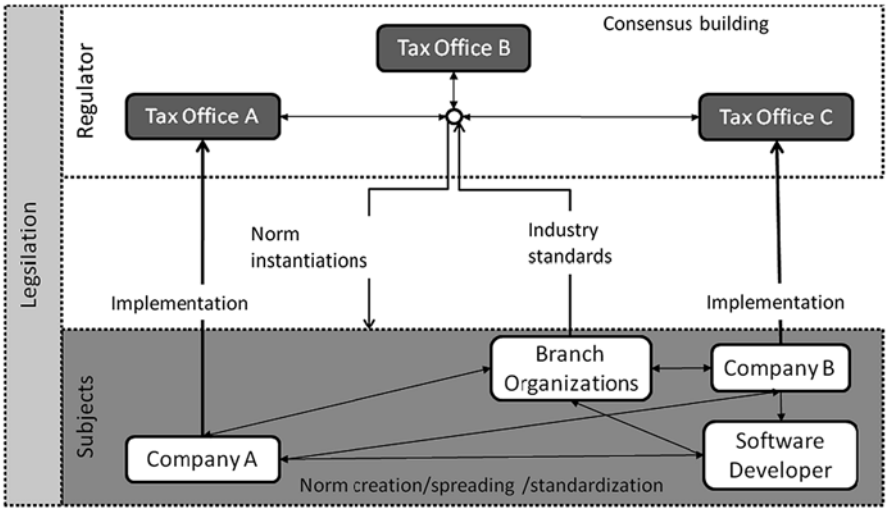

Fig. 7.1 Bounded norm emergence in the Dutch taxation approach

approach of other companies. These implementations and technologies are then proposed to Dutch tax. Dutch tax guides the norm emergence process by approving or disapproving the proposed implementations as instantiations of the legislative norms. Different regional offices of Dutch tax also have to agree which implementations are acceptable to ensure a fair compliance assessment. We consider norms to be emerged, when the proposed implementations are approved by the tax administration. Approved implementations may function as benchmarks for future audits or for future implementations of the guidelines (Gribnau, 2008).

Data Collection

Data for this case study was collected by the following methods: document analysis and semi-structured interviews. We studied public documents from the Dutch Tax Administration and legal judgments on disputes related to the kilometer administration. Furthermore we conducted two interviews with experts from Dutch tax on the use of new technologies for kilometer registration. One expert has a background in both chartered certified accounting and IT auditing; the other expert was more into legal aspects of taxation. Notes were made of both meetings and verified later by the interviewee.

In the next section, we study two scenarios of norm emergence. One scenario is brought about by the introduction of new technology, and second concerns

context-specific problems with the implementation of legislation. For both cases we describe the spreading of the implementations of the norms as well as the consensus building about the acceptability of those implementations among tax auditors, branch organizations, technology providers, and companies. Furthermore, we describe the differences between the old norm and the newly emerged norm.

Scenario 1: The Use of New Technology

Nowadays technologies such as navigation software, GPS, and black box systems are frequently implemented in cars to monitor and log all kinds of events and provide the drivers with better services. In the case of GPS-based navigation software, GPS coordinates can be recorded and kilometers driven can be calculated for trips planned with the routing software. Advantages of the navigation software over the odometer recordings are the following: distances driven can be directly coupled with location and (when supported) with time information and human error-prone actions can be limited as registration and generation of administrative report may occur automatically.

Lease car drivers more often propose registrations, made by the navigation software, as a means for their kilometer administration to tax officials. Using navigation software for administrative purposes thus seems to be already spreading as a (socially accepted) norm. However, the regulations prescribe that the norm is that a kilometer administration should be based on the recordings of the odometer. Currently, tax officials thus cannot accept kilometer administrations based on recordings made by the navigation software.

What are the functionalities and (legal) requirements that should be met, before registrations made by navigation software can count as a replacement for the kilometer administration based on an odometer? The regulations prescribe two obligatory legal aspects: the kilometer administration should be sound (complete and valid) and auditable. In the table below we describe for both the odometer and navigation software which evidence is gathered to prove one complies with the legal aspects (Table 7.1).

In the table we see that both systems rely on similar and different approaches to comply with the legislation. Similarities can be found in the additional controls to determine the validity of the administration. Differences can be found in the amount of human vs. automatic actions and the type of controls that are used: physical vs. digital controls.

In general, the replacement of human actions for automatic processes increases the soundness and auditability. Human errors (wrong or forgotten registrations) are reduced, and controls can be built in IT systems to enable logging and monitoring for an (complete) audit trail. In addition, the automated route registration limits the possibility for lease car drivers to manipulate the data for their advantage.

Table 7.1 Approach to compliance for the odometer vs. the navigation software

			Odometer	Navigation software
Sound	Completeness	Measurement	Recorded distance for a certain period	Recorded trips
		(Additional) Controls	Direct observation Garage maintenance checks	Logging user actions System configuration
	Validity	Measurement	Manual registration of trips	Automatic logging GPS coordinates
		(Additional) Controls	Working schedules Traffic cameras, fines	GPS coordinates of destinations Working schedules Traffic cameras, fines
Auditable	Data retained		Physical recording device (Manual) Calculations	Digital recording Automatic

Furthermore, tax officials can directly retrieve the source data from the navigation system and link them to claims made in the fiscal report.

Different types of controls have different strengths and weaknesses. The recordings of the odometer can only be directly observed but are quite reliable as an odometer rarely breaks down. Tampering with the odometer requires some reasonable mechanical expertise. Navigation technology on the other hand can easily facilitate tampering with the data when retroactive registration or changes to the logged data are options included in the software. Furthermore, navigation systems based on GPS technology may suffer from breakdowns and continuity issues through blind spots (locations may be out of reach for GPS transmitters). These problems strongly affect the soundness and auditability requirements, but they can be overcome through logging and a complete audit trail. Thus, before navigation software can be used as a replacement for the odometer, it needs controls to ensure that the data is secured, can be retained, and is complete. We summarize the steps and interactions of the norm emergence process in the figure below.

In Fig. 7.2 we see that tax offices have agreed upon the use of an odometer to register kilometers. Companies and software developers have come up with the navigation software as a means to register kilometers. Company B proposes to tax office C the use of navigation software. The tax offices agree that the use of navigation software for a kilometer administration will only be allowed in the future (dotted arrow) when the software meets certain (legal) requirements. They respond to the company and software developer to come up with software that complies with the legislation. The exact implementation stays as the responsibility of company and software developer.

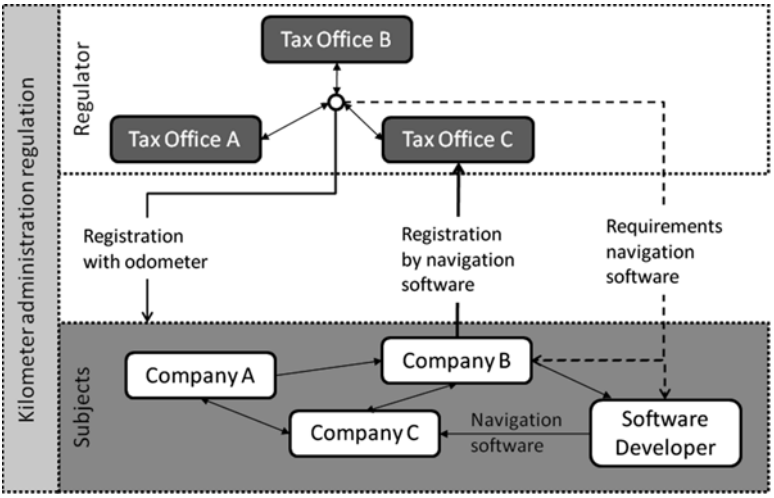

Fig. 7.2 Norm emergence through the proposition of navigation software as a means to comply with the kilometer administration regulations

Scenario 2: Context-Specific Implementation Problems

The tax administration together with (interest) groups and branch organizations EVO, VNO-NCW, Uneto-Vni, Bouwend Nederland, and Fosag has concluded an agreement in 2009 to allow a simplified kilometer registration for delivery vans (Belastingdienst, 2009). When, due to the activities of a company, delivery vans are used for many of the trips, keeping a kilometer registration can become quite an administrative and financial burden for both employee and employer. Employee and employer can then decide in a written agreement that the employee makes use of this new amendment. The amendment prescribes that instead of keeping a detailed registration of each business and private trip, an employee using a delivery van can prove to the tax administration the total number of driven private kilometers using a combination of:

A simplified kilometer registration kept by the employee
The business addresses recorded in the (project) administration of the employer

When the simplified kilometer registration regulation is applied, private use of the vehicle during work and lunch time is not allowed. For example, commuting from work to home during lunch time will count as private use of a lease car. Note that having a kilometer registration remains obligatory, but an employee does not need to include all the detailed information on single trips as long as the required information can be obtained from the accounts of the employer. Besides the brand, type, and license plate number, an employee should record daily in its simplified registration:

Date
Working hours
Start and end position of the odometer

When the employee uses the vehicle for a private trip after working hours he or she should also record:

Date
Start and end position of the odometer for the private trip
Departure and arrival address

The employee does not have to register the business trips because the (project) administration of the employer should contain the sequence of the business addresses visited in 1 day. The project administration of the employer therefore should contain:

The business addresses that an employee visits on a certain day
The sequence in which the addresses are visited

When the employer does not register the sequence in which the addresses are visited, the employee should record the sequence of business trips in its simplified kilometer registration and refer to the addresses in the (project) administration. The employee uses, for example, the project name or the account number that is also used in the company's administration.

If the employer notices a difference between the number of business kilometers in the simplified kilometer administration and the number of business kilometers in the (project) administration, an explanation for the difference is required. For example, differences can occur through detours due to roadwork or the licensed purchase of supplies. In the latter case the employer can use bills in the administration as supporting evidence.

Compared to the original implementation, the data that must be registered does not change. When all relevant information sources are combined, the information of individual trips of individual employees can still be retrieved. What does change in the new situation is the introduction of a shared responsibility of employee and employer to guarantee the accuracy and auditability of the kilometer registration. It is necessary that the responsibilities of both employer and employee are clarified upfront in an (written) agreement. After all, both parties depend on the quality of data of the other. We summarize the steps and interactions of the norm emergence process in the figure below.

In Fig. 7.3 we see that the standard registration (of each individual trip) causes difficulties in companies where employees use delivery vans. The companies mention their problems to branch organizations and interest groups. Together these organizations make the problems explicit to the tax administration. The individual tax offices agree upon the use of a simplified kilometer administration, when employee and employer together can ensure that the data to comply with the regulations is available. The tax administration provides feedback on the adapted regulations to all subjects in the environment.

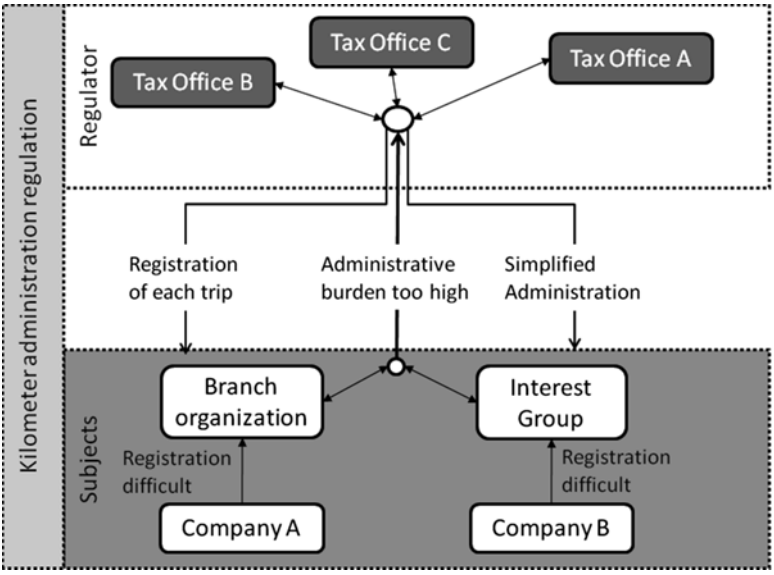

Fig. 7.3 Norm emergence through implementation problems in kilometer administration for delivery vans

Discussion

In the case study we describe two specific scenarios of norm emergence for the regulation that "an employee should keep a sound and auditable kilometer administration; to prove that one does not drive over 500 private kilometers on a yearly basis," Article 13bis (LB, 1994). The first scenario describes norm emergence through the introduction of new technology. The use of navigation technology was spread among technology providers and companies in the environment. A company then proposed to a tax office the use of navigation software instead of the odometer to comply with Article 13bis. These processes correspond to the two variations, norm spreading and bottom-up norm creation, described by Savarimuthu and Cranefield (2009). However, they do not describe the next process step where the tax administration restricts the emergence of that norm. The tax offices together agree that navigation software as a means to register kilometers is only allowed when it meets certain technical requirements that follow from the legislation. These requirements are then instantiated in a new amendment issued by the government.

The specification of abstract norms into concrete implementations has been well described by Dignum (2002) and Aldewereld et al. (2006). Their approaches only discuss specification of norms from the viewpoint of the institution: norms are specified in terms of the ontology (Dignum, 2002) or the concrete concepts (Aldewereld et al., 2006) used by the institution. Their approaches do not discuss the fact that norms may be specified by the individual agents in terms of their

internal mechanisms or architecture. Compliance checking of norms that emerged bottom-up is therefore also not discussed. Another related issue to the restriction of norm emergence is that it brings about additional goals for the subjected agent. Companies and software developers that want to motivate the use of navigation software as a means of kilometer administration must develop new controls that need to be embedded in the software.

A simplification of reality used in the literature is that an institution is seen as a single agent. However, institutions like the Dutch Tax Administration in our case in fact consist of multiple agents and departments, for instance expert groups with different backgrounds (technological, legal) or regional tax offices. There may be many disagreements among departments, but to the outside world the tax office has to present a unified view. In a way, such a unified view must be enacted through organizational norms. We therefore hypothesize that at the institutional level similar processes occur as the processes that occur among the interactions of individual agents that are subject to the regulations. Here too, open norms must be interpreted and enforcement must be operationalized.

The second scenario describes norm emergence that occurred through problems with the implementation of a kilometer registration among delivery vans. Employees and companies that use their lease vehicles for multiple (short) business trips per day suffered from an administrative and financial burden when registering all individual trips according to the regulations. Furthermore, information was registered twice by the employee and (project) administration of the company. Interest groups, branch organizations, and the tax administration concluded an agreement that a simplified kilometer administration was allowed for delivery vans, when employee and employer together could provide the information required by the regulations.

These processes can be described in terms of mechanisms found in the literature (Sect. 2). We see that representatives of agents with a certain position in the community (interest groups) fulfill a key role in the emergence of a simplified kilometer administration. Besides that, we see interaction of the events occurring at the micro and macro level. Macro properties (legally required elements of the kilometer registration) affect the functioning of individual agents (record business and private trips). Micro-social behavior (visiting multiple business addresses per day, keeping a project administration, recognizing administrative burden, practical inability to keep a reliable records) resulted in social interaction (complaints among agents, talks between branch organizations and the tax office). Finally, this resulted in new macro-social properties (simplified administration for delivery vans).

A similar situation can also arise in an artificial agent community, when an agent simply cannot comply with the norms due to shortcomings in its architecture. In that case, is the agent violating the norm and should it be punished or should another response follow? After all, it should be possible at least to comply. Furthermore, this issue is also of interest in the light of literature on normative beliefs and emergence (Conte & Castelfranchi, 1999). When the agent believes that the norm is obligatory, but cannot prescribe to it, what would be then the agent's contribution (positive or negative) to the spreading of the norm?

In the case of the kilometer registration, the solution to the problem of incompliance was to share normative responsibilities. Information collected and kept by employee and employer together does make compliance of the employee possible. In MAS research coalitions are described as a means to perform actions and to reach goals that can be reached by individual agents. To our knowledge shared compliance has not been a topic of coalition studies in MAS research.

Conclusions and Future Research

Research on social simulations in the MAS field has formed theories that try to explain or reproduce norm emergence. Often these theories choose either a top-down and prescriptive viewpoint or a bottom-up and social convention viewpoint. We consider a hybrid form: bounded norm emergence. Here, the bottom-up emergence of norms is restricted by the limits of relatively abstract norms, which are given top-down. Bounded norm emergence typically occurs over time when open norms are implemented by subject agents and approved or enforced by regulator agents. To explain the phenomenon of bounded norm emergence, a combination of both perspectives is needed. In this chapter we propose to use techniques taken from simulation studies in MAS, in particular about norm emergence and norm distribution, and apply them in a setting of top-down norm prescription, where the norms are relatively abstract or open for interpretation.

A major shortcoming of the bottom-up approach to norm emergence is that issues of compliance and enforcement are not addressed. Unlike such frameworks, which allow basically any behavior to emerge as a norm, in our case study the emergent norms are constrained by an abstract norm, which needs to be interpreted before it can be implemented and audited. A related issue is that when norms emerge they may also bring about new goals for the agents. In the case study we discuss a general norm—one must keep a kilometer registration—which allows many technical implementations. To approve these technological innovations and consider them valid instantiations, new norms must be established and the legislation may have to be adapted to allow a fair assessment of compliance.

By contrast, most top-down approaches have an institution-centered perspective to norm specification. Abstract norms are translated to concepts defined by the institution, rather than concepts defined by the subjects. Therefore, some norms are simply impossible to comply with. In an artificial agent society, for example, a norm which cannot be expressed in terms of the architecture of the agent is unattainable. In the case study, keeping a kilometer registration of a company van turns out to be practically impossible. Only by combining information from employer and employee the administrative burden can be lessened.

Traditionally, regulators can only approve or disapprove certain implementations of an abstract norm. In our combined approach, we can also account for the fact that regulators have an interest in simplifying norm enforcement practices. In the case study, we describe the process of companies and interest groups proposing norm implementations to the tax administration, who react by establishing technical standards.

Furthermore, an interesting observation is that group processes also occur at the institutional level. In the case study all regional tax offices and expert groups with the tax department have to agree before norm implementations can be approved and emerge as a norm. We hypothesize that bottom-up theories about norm emergence and distribution can also be applicable in this case.

References

Aldewereld, H., Dignum, F., Camino, A. G., Noriega, P., Rodriguez-aguilar, J. A., and Sierra, C. (2006). *Operationalisation of norms for usage in electronic institutions*. Paper presented at the Proceedings of AAMAS.

Ayres, I., & Braithwaite, J. (1992). *Responsive regulation: Transcending the deregulation debate*. Oxford: Oxford University Press.

Belastingdienst. (2009). Vereenvoudigde rittenregistratie voor bestelauto's (Simplified Trip Registration for Delivery vans): Belastingdienst Available online at http://www.belastingdienst.nl/wps/wcm/connect/bldcontentnl/belastingdienst/prive/auto_en_vervoer/u_reist_naar_uw_werk/auto_van_uw_werkgever/afwijkende_regels/vereenvoudigde_rittenregistratie_voor_bestelautos.

Braithwaite, V. (2007). Responsive regulation and taxation: Introduction. *Law and Policy, 29*(1), 3–10.

Burgemeestre, B., Hulstijn, J., and Tan, Y.-H. (2009a). *Agent architectures for compliance.* Paper presented at the Proceedings of the 10th Annual International Workshop Engineering Societies in the Agents World.

Burgemeestre, B., Hulstijn, J., and Tan, Y.-H. (2009b). *Rule-based versus principle-based regulatory compliance*. Paper presented at the JURIX'09.

Conte, R., Andrighetto, G., Campenni, M., and Paolucci, M. (2007). *Emergent and immergent effects in complex social systems*. Paper presented at the Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence.

Conte, R., & Castelfranchi, C. (1999). From conventions to prescriptions: Towards an integrated view of norms. *Journal Artificial Intelligence and Law, 7*(4), 323–340.

Dignum, F. (2002). *Abstract norms and electronic institutions*. Paper presented at the Proceedings of the International Workshop on Regulated Agent-Based Social Systems: Theories and Applications (RASTA '02).

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review, 14*(4), 532–550.

Ford, C. L. (2008). New governance, compliance, and principles-based securities regulation. *American Business Law Journal, 45*(1), 1–60.

Goldspink, C. (2009). Social self regulation in on-line communities: The case of wikipedia international. *Journal of Agent Technologies and Systems, 1*(1), 19–33.

Gribnau, H. (2008). Soft law and taxation: The case of the Netherlands. *Legisprudence, 1*(3).

Wet op de Loonbelasting 1964 (Income tax act) (1994).

Sadiq, S. W., Governatori, G., and Namiri, K. (2007). *Modeling control objectives for business process compliance.* Paper presented at the Business Process Management (BPM 2007).

Savarimuthu, B. T. R., and Cranefield, S. (2009). *A categorization of simulation works on norms*. Paper presented at the Dagstuhl Seminar Proceedings 09121: Normative Multi-Agent Systems.

Vazquez-Salceda, J., Aldewereld, H., & Dignum, F. (2005). Norms in multiagent systems: From theory to practice. *International Journal of Computer Systems Science & Engineering, 20*(4), 225–236.

Yin, R. K. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks: Sage Publications Inc.

Chapter 8
Norm Dynamics Within the Mind

Giulia Andrighetto, Daniel Villatoro, and Rosaria Conte

Social norms are largely regarded as solutions to the problem of attaining and maintaining social order (Axelrod, 1986; Durkheim, 1950 [1895]; Fehr & Fishbacher, 2004; Posner, 2000). It is argued that the norm of reciprocity, for example, solves what is currently known as the *puzzle* of human cooperation (Axelrod, 1986; Boyd & Richerson, 1988; Gintis, 2003), revolving around the following question: How can self-defeating behaviour, like giving help, compete with self-enhancing strategies, like not reciprocating the received help, and successfully spread over a given population?

The answers mainly focus on the *types* of interaction strategies and their distribution among sub-populations. Undoubtedly, if only unconditional strategies exist (e.g. cooperation and defect), defectors will outcompete cooperators. But if we add a conditional strategy, for example a *tit for tat* strategy, players following it will soon outcompete the unconditional defectors (Axelrod, 1997). Analogously, if a group of agents responsible for punishing defectors joins the population (composed only of unconditional cooperators and defectors), it will lead to a reduction in exploitation and consequently to the survival of cooperators (Bowles & Gintis, 2004).

Within the approach to norm dynamics described so far, the social process investigated is a *one-way* process. As reported by Axelrod (1997), within evolutionary game theory social actors are *taken for granted*. Novelty can only emerge from the bottom up. New strategies are accounted for only as mutations that will be selected positively if advantageous and discarded if disadvantageous. To remedy this shortcoming,

G. Andrighetto (✉)
LABSS-ISTC/CNR, Rome, Italy

European University Institute, Fiesole, Italy
e-mail: giulia.andrighetto@gmail.com

D. Villatoro
IIIA-CSIC, Barcelona, Spain

R. Conte
LABSS-ISTC/CNR, Rome, Italy

M. Xenitidou and B. Edmonds (eds.), *The Complexity of Social Norms*,
Computational Social Sciences, DOI 10.1007/978-3-319-05308-0_8,

Axelrod proposed a "tribute" model (1997) in which social actors grow by means of extortion: stronger agents are supposed to extort tributes from the weaker and eat them up if these refuse to pay. Via this Mafia-type affiliation, the model allows actors to emerge at supra-individual levels of aggregation.

While the tribute model shows how coalitions are formed, it does not show how individual actors, i.e. their behavioural rules, change. In the evolutionary approach to positive social action it is not individual actors that change, but lineages. In learning models of cooperation and coordination, individual strategies change under winners-stay-losers-change type of rules or under the effect of imitation. However, or the rules or the other mechanisms governing behaviour are rigid, changing essentially by means of reproduction.

We claim that not only behaviours but also behavioural rules change, either by means of evolutionary processes or by learning. They may be reinforced under the perceived effects of previous actions, a process known as second-order emergence (Gilbert, 2002). But they may also undergo a more radical influence. Under the effect of social perceptions, agents may develop new rules for action, new beliefs, goals and expectations and even new decision mechanisms. As we shall see, this is the case with social norms.

In this chapter, we discuss the necessity for a theory of norm dynamics that accounts for a process occurring not only at the observable, *behavioural* level, but also at the non-observable, *mental*, level and illustrate the work done in this direction by the authors. We represent norm dynamics as an *iceberg*, with the water line standing for the boundary between the observable domain, behaviour, and the unobservable one, the mind. As usual, the portion of the iceberg below the observable level is much larger than its tip, signifying that most of the processing occurs in the mind, below the observation line. Thus, to properly model norm dynamics, this hidden part requires to be carefully explored. When emerging, spreading, innovating and declining, the process that norms undergo is a *complex bidirectional* one, which includes the way up and the way down, and consists of the interplay between the social and mental dynamics (see section "Mental and Social Dynamics of Norms").

In particular, in this work we focus on two specific dynamics: norm *emergence* and norm *internalisation*. First, we will ask how a norm can emerge. In particular, we claim that norms emerge in society *by* and *while* shaping and modifying the representations and mental mechanisms of the society's members, i.e. while *immerging* in their minds (Andrighetto, Campennì, Cecconi, & Conte, 2010; Andrighetto, Campennì, Conte, & Paolucci, 2007; Castelfranchi, 1998; Conte, Andrighetto, & Campennì, 2014; Conte, Andrighetto, Campennì, & Paolucci, 2007). Since norms require a set of corresponding mental representations to support them, we will provide an explanation of how these arise. Second, we will investigate the conditions under which compliance to a norm becomes independent of external enforcement, i.e. when the norm addressee observes it free from external punishment or rewards.

To account for the complex dynamics of norms, a simple agent model is insufficient. Agents must be endowed with a considerable mental capacity, enabling them to represent norms and accomplish a number of mental operations on these representations

(see section "Normative Architecture EMIL-I-A"). To show why, we will refer to two simulation-based studies of norm dynamics that we have recently carried out. In the first study (reported in section "Simulating Norm Emergence: Behavioural Contagion vs. Norm Immergence"), we investigated the conditions under which norms emerge in a multi-setting world, comparing simple agents (conformers) with more complex (normative) cognitive agents. Results will be argued to show that unlike conformers, i.e. agents following a mere imitation rule, normative agents are able to converge on the same behaviour when moving from one context to another. In the second study (reported in section "Simulating the Effect of Norm Internalisation in Promoting Cooperation"), we looked at norm dynamics in a population of normative agents enabled to internalise norms, i.e. to learn to comply with norms even independent of external sanctions. We have compared the level of cooperation obtained by agents enabled to internalise norms with that of agents that do not have this capability. Results show that the former type of agent cooperates much more than latter.

Mental and Social Dynamics of Norms

Unlike a pure behavioural account of norms, the proposed approach aims to explain norm's compliance based on mental representations, i.e. normative beliefs, goals and expectations.[1] As proposed by Ullman-Margalit (1977), we refer to social norms as *prescribed* guides for conduct informally transmitted from one agent to another through normative requests of the type "one must keep to one's commitments" and "you should not ask what your country can do for you, but what you can do for your country" and sometimes conveyed under evaluations in the form "smoking is antisocial behaviour."[2] It is their prescriptive strength or, to state it with Gilbert (1983), *mandatory* force that makes norms differ from mere social habits. In order to motivate people to comply with them, social norms and their prescriptive character need to immerge into people's minds and shape their mental representations (Andrighetto et al., 2007; Andrighetto, Campennì, et al., 2010; Castelfranchi, 1998; Conte et al., 2007, 2013).

A similar approach to social norms has been provided by Bicchieri (2006). She presents her model of social norms as a rational reconstruction of the conditions under which social norms can be taken to guide action. According to Bicchieri, two conditions must be satisfied for a social norm to exist in a given population. First, a sufficient number of individuals must know that the norm exists and applies to a situation. Second, a sufficient number of individuals must have a *conditional preference* to comply with

[1] For an analysis of the cognitive requirements of norm compliance, see also Xenitidou and Elsenbroich (2010).

[2] Although linguistic communication is a very effective means for transmitting normative requests, a large amount of information about how one should behave in a certain situation should be inferred through the observation of others' conduct.

the norm, given that the right expectations are satisfied. Bicchieri distinguishes two types of expectations that must be satisfied for conditional compliance with social norms to be obtained. By *empirical expectations*, Bicchieri refers to the *belief* that enough other people in a similar situation obey the norm (or have done so in the past). By *normative expectations*, she means the *belief* that enough other people *think* we ought to obey the norm in that situation and may be willing to sanction us in a positive or a negative way depending on our decision to comply or not with the norm. Norm compliance is conditional on the (empirical) expectation that a sufficient number of people conform to the rule and on the (normative) expectation that other people expect her to follow the rule as well and possibly enact positive or negative sanction for transgression/conformity.

Though extremely interesting, the approach proposed by Bicchieri does not satisfactorily explain the mental process allowing normative and empirical expectations (i.e. normative knowledge) to motivate autonomous agents to comply with norms. In other words, it is not entirely clear how the belief that enough people in a similar situation obey the norm and the belief that enough other people think we ought to obey the norm in that situation (and may even be willing to sanction us in a positive or a negative way depending on our choice to obey or not the norm) may motivate people to comply with norms.

In this section, we describe a multilevel model of norm immergence aimed to account for the mental path followed by a norm in regulating human behaviour and more specifically to shed light on the cognitive ingredients and processes necessary for a normative request to be complied with.

Norm Immergence at the Epistemic Level

In a view of norms as two-sided, external (social) and internal (mental) objects (Conte & Castelfranchi, 1995, 1999, 2006), social norms come into existence only when they emerge, not only *through* the minds of the agents involved but also *into* their minds. We claim that a norm emerges only when the associated normative belief immerges in the minds of the agents and the corresponding normative goal and expectations are formed and pursued.[3] This result is usually generated through a number of intermediate loops. Before any global effect emerges, specific local events affect the generating systems (e.g. agents), their beliefs, goals and operating rules in such a way that agents are more likely to reproduce the macroscopic effect.

The first step for a norm to immerge is its recognition by the norm addressee. Exposed to the normative behaviours of others and to their explicit or implicit normative

[3] We refer to beliefs and goals as internal representations triggering and guiding action: beliefs represent the current state of the world, while goals represent the state of the world that agents want to reach by means of action and that they monitor while executing the action. In general, goals are a subset of the motives or the reasons for action that can be generated, updated and dropped (Conte & Castelfranchi, 1995) under the effects of new beliefs about changing circumstances.

requests, agents possibly acquire normative belief. More specifically, any agent y recognizing a given input as a norm forms at least the first of the following beliefs:

1. Main normative belief (indicating the existence of the norm), which states that a given type of behavior B, in a particular context C, for a given set of agents S, is forbidden, obligatory, or permitted. More precisely, the belief states that "there is a norm N prohibiting, prescribing, or permitting a." Beliefs supporting the creation of main normative beliefs include:
   - The source of the prescription is a formal authority, held to issue (a specific set of) norms.
   - The source is not a formal authority, but the set of agents S, which y belongs to.
   - The source is a distributed one.
   - N is impersonally addressed; i.e., anyone belonging to S in circumstances C is required to comply with N.
2. Normative belief of pertinence (indicating that the belief's holder belongs to the set of agents on which the norm is impinging): y believes she belongs to S.
   With these necessary normative beliefs, one more is often, but not necessarily, associated:
3. Norm-enforcement belief: the belief that normative compliance and violation are supported or enforced by positive or negative (informal) sanctions.

With their conduct, individuals communicate not only that there is a norm governing a certain situation but also that they *want* and (explicitly or implicitly) *ask* that others comply with it. The normative actions of others communicate both that there is a norm regulating a certain situation and that there is a widespread request for it to be fulfilled.

Norms are influencing devices that require the altering of the goals of the individuals subject to them. The notion of normative expectation, as proposed by Bicchieri (2006), does not explicitly imply an individual (or a set of individuals) acting to modify somebody else's goals, while this *influencing goal* characterises the normative request. Interpreting an action as a normative request implies recognising that there is somebody asking you to adopt his or her goal, i.e. to comply with the norm. The normative request also presupposes that you are asked to comply with the norm not because of the personal goal of the requester, but because norms must be complied with. The recognition of a widespread request for compliance allows the norm immergence process to start and, as we discuss later in this chapter, the activation and adoption of normative goals to take place.

Moreover, the normative actions of others are also important cues through which how salient the norm is can be inferred. We refer to salience as the measure indicating how much a norm is prominent within a group and a given context (Andrighetto, Villatoro, & Conte, 2010; Andrighetto et al. 2013; Bicchieri, 2006; Cialdini, Kallgren, & Reno, 1991; Houser & Xiao, 2010). The amount of compliance (Cialdini et al., 1991), the surveillance rate, the probability and intensity of punishment, the enforcement typology (private or public, second and third party, punishment or sanction, etc.) (Galbiati & Vertova, 2008; Houser & Xiao, 2010; Masclet, Noussair, Tucker, & Villeval, 2003), the efforts and costs sustained in educating the population

to form a certain norm, the visibility and explicitness of the norm and the credibility and legitimacy of the normative source (Faillo, Grieco, & Zarri, 2013; Villatoro et al. in press) are all signs through which people infer how important and active a social norm is in a specific context. The more salient a norm is considered, the higher the probability to be complied with.

However, recognising that there is a prescription and a *widespread* request for it to be fulfilled and that its compliance is enforced through positive and negative sanctions are necessary but insufficient conditions for compliance with norms. The norm immergence process should go further and also influence the *motivational side*, modifying individuals' goals.

Norm Immergence at the Motivational Level

Unless we consider norm compliance as an automatic reactive process, a normative belief must give rise to a normative goal for the subject to act in accordance with the norm itself. This process is more or elaborated and governed by different mechanisms. From the least sophisticated, where norm obedience becomes automatic, leaving little room for autonomy, to the most complex, such as instrumental norm adoption, i.e. the calculation of the advantages and disadvantages of norm compliance, a mechanism enabling an external command to become a goal is needed. We refer to the mechanism that leads from a normative belief to the generation and adoption of a normative goal (i.e. norm adoption) as *norm adoption mechanism* (see Conte & Castelfranchi, 1995).

An autonomous cognitive agent acts always for his or her own final motives and purposes and has to have reasons for choosing to act as he or she does. Thus, an agent (the adopter) will adopt another agent's goal (i.e. the adoptee's goal) as his or hers, on condition that he or she, the adopter, comes to believe that the achievement of the adoptee's goal will increase the chances that he or she will in turn achieve one of his or her previous goals.

An agent can decide to adopt a normative goal for several higher motives (for a detailed analysis, see Conte & Castelfranchi, 1995):

- *Instrumental* motives: The subject adopts the normative goal if he or she believes that he or she can get something in return (avoid punishment, obtain approval, praise, etc.).
- Cooperative adoption is a particular form of instrumental adoption, in which the subject adopts the normative goal to achieve, not a personal, but a common goal. Norm-adoption is cooperative when it is value-driven; that is, when the subject shares both the goal of the norm and the belief that the norm achieves it. For example, an agent may decide to conform to the recycling norm because he believes that, by doing so, he helps reduce our species' negative impact on the environment.
- *Terminal* motives: The subject wants to observe the whole set of norms addressing his or her as ends in themselves. He or she has the terminal goal or value that "norms should be respected" (Kantian morality). Terminal norm adoption implies that any norm deserves obedience until it exists.

Normative goals can be formed for different reasons, also for self-regarding reasons, as in instrumental norm adoption. This does not prevent the goal thus formed from being normative in the fullest sense: a normative goal is a goal relativised to a normative belief, held because and to the extent that it is believed to be exacted by a norm. All that is needed for a goal to be normative is that it is based on norm-related representations.

Thus far, we have modelled how and when we generate a goal relativised to a normative belief. Is this enough for the norm to be actually observed? Can we say that such a condition is sufficient for norm compliance? Unfortunately not. The way to normative action is still quite long and interspersed with check points in which decisions might endanger the whole process. The normative goal may be dropped at any point, along this complex itinerary. Worse, it may be the case that the normative goal is never dropped but the norm is not complied with. This is the case when interferences are beyond one's control. If the seat belt of my car breaks up while I am sitting in a traffic jam, there is little I can do but violate the norm. But there are other check points. First of all the new goal is checked against the current state of the world. Over time, things may have changed. The goal might turn to be already true in the world. Interferences might have gone or been superseded by other events. A second check consists of evaluating the goal against other goals. If it is found incompatible with other more important ones (normative or non-normative), it will probably be dropped.

The Complex Loop of Norm Emergence and Immergence and Its Behavioural Consequences

Why care about the reasons for conformity once convergence has emerged one way or the other? There are several answers to this question. One for example is that we are interested in predicting not only degrees of compliance but also other behavioural consequences of norm emergence. If people are motivated by a norm to converge on a particular behaviour, and not by mere imitation or social conformity, they will be more willing to defend and enforce it. For example, they might be willing to send out educational messages and impose sanctions or punishments on norm violations. The more a behaviour is believed to be prescribed and to be a widespread goal of the entire group, the more it will be complied with, and the more, in turn, the corresponding prescription will be enforced. Therefore, the spread of normative influence contributes to the spread of normative beliefs and vice versa. This complex loop guarantees the stability and robustness of the emerged process and possibly leads to the internalisation of the norm (Andrighetto & Villatoro, 2011; Conte, Andrighetto, & Campennì, 2014; Conte & Dignum, 2001; Villatoro, Andrighetto, Sabater-Mir, & Conte, 2011).

As claimed in Andrighetto, Villatoro, et al. (2010) internalisation occurs when norm's compliance becomes independent of external enforcement, i.e. when the norm addressee observes it free from external punishment and reward. In other words, it is a mental process that takes a (social) norm as input and provides the individual with *terminal* goals, i.e. goals that are considered as *ends* in themselves

instead of means for achieving other goals. This process has several advantages. For example, norm compliance is expected to be *more robust* when norms are internalised than in cases where norms are external reasons for conduct: if everybody in the population internalises a norm, there is no incentive to defect and the norm remains stable (Gintis, 2003). Driven by terminal motivations, individuals who internalise a norm are much better at not only complying with norms but also defending them than are externally enforced individuals (see section "Simulating the Effect of Norm Internalisation in Promoting Cooperation"). An effect of the latter prediction is that norm internalisation is decisive, if not indispensable, for *distributed* social control. Internalisation is not only a mechanism of private compliance but also a key factor of social enforcement. Individuals who have internalised the norm comply with it with no need for external enforcement, and in many circumstances also want to persuade others to observe the norm, by reproaching transgressors and reminding would-be violators that they are doing something wrong.

Norms may sink at different levels below the iceberg's water line. At the first level, immergence generates normative states and operations and agents still need some reasons for complying with them. But norm compliance is not always *deliberative*. As norms are plunged into the mind their external normative origin gradually gets lost. Once internalised norms start to operate as fully endogenous goals, until they become integrated with action plans or become part of sensory–motor responses triggered by given stimuli. At this point, norms are complied with *thoughtlessly* (Epstein, 2007), and the norm-related actions become semi-automated routines. However, internalised norms are not bound to remain such: depending on circumstances, agents may retrieve awareness of their exogenous source and of their external enforcement (for a detailed description of the norm internalisation process, see Andrighetto, Villatoro, et al. 2010). Norms' dynamics both outside and inside the agents' mind is a continuous and multilevel process.

Simulating Norm Emergence: Behavioural Contagion vs. Norm Immergence

Recent simulation data (Andrighetto, Campennì, et al., 2010; Campennì, Andrighetto, Cecconi, & Conte, 2009; Conte, Andrighetto, & Campennì, 2014) show that under specified conditions mere imitation is not sufficient for achieving convergence and immergence is required. One of the structural conditions under which conformism barely yields convergence is the *multi-setting* world, i.e. a world in which agents move among settings based on personal sequences and linger on each of them according to personal agendas. When living in a multi-setting world and continuously moving among different social contexts, agents acting only through behavioural contagion and passive social impact are unable to converge on a single specific action, while normative agents are. When they have to move across different settings, agents endowed with the ability to recognise social norms and to generate the corresponding mental representations use the normative representations as a

device providing instruction about the action on which to converge. Once norms have immerged in their minds, for deciding how to act, agents do not need to constantly monitor what others are doing. They are less dependent on contingencies and less prompt to abandon their normative conduct in order to follow what others do.

In real life, agents move from public offices to private residences, from sport and shopping centres to underground stations and from these to cinemas, pubs, etc. Suppose that different options for action are available in each setting. For example, you can play music, eat and drink in pubs; get undressed, work on your biceps and take a shower in a fitness centre; buy a ticket, take a seat and watch a movie at the cinema, etc. Suppose also that there is one action common to all settings—say, *joining a queue*, if there is one, at each entrance. Since they continuously move from one setting to another, how can simple conformers interpret the common action of joining a queue as normative and converge on it?

Our simulation data (Andrighetto, Campennì, et al., 2010; Campennì et al., 2009) show that even conformers with a persistent memory take long to converge, thus not allowing the social norm to emerge. Instead, agents converge more easily and faster when enabled to form normative mental representations and act based upon them. Furthermore, in a multi-setting world, norm immergence produces a different observable dynamic than other simpler rules do. In particular, as to *within*-setting comparison, norm immergence yields a fuzzier distribution than conformism: some normative agents form and adopt different norms in the same setting, while conformers rapidly converge on the same action (due to the fact that their behaviour is strongly influenced by neighbours). But whilst *between*-setting distribution presents sharp boundaries among conformers, who barely converge on the common action, it is much smoother among normative agents, who are more autonomous and tolerate perturbation but gradually converge on the common action. Normative agents converge while preserving their autonomy: they choose how to act considering the normative beliefs they have formed while observing and interacting with others. Thus, they converge in a more stable way: after a certain period of time, the majority of agents start to perform the same action. It is possible to say that a norm has emerged after having immerged into their minds.

To fully operationalise such dynamics between norm immergence and emergence complex agent architecture is required. In the rest of this chapter, one such type of agent architecture, i.e. EMIL-I-A, is presented. EMIL-I-A is endowed with mechanisms allowing norms to immerge into the agents' mind at various levels of depth, up to the deepest, i.e. norm internalisation.

Finally, some simulation results showing how norm internalisation promotes cooperation within groups facing a social dilemma are presented and discussed.

Normative Architecture: EMIL-I-A

Operational models of the *multidirectional* dynamics of norms are still lacking. For the most part, existing work concentrates on the way up of the process leading to norms being established, i.e. on the behavioural interaction (Chakrabarti & Basu, 2010;

Sen & Airiau, 2007; Young, 1993, etc.). The complementary side of the process, i.e. the way back from the macro- to the micro-level, is poorly investigated and little implemented. To operationalise the norm immergence process a complex agent architecture is required endowed with the capability of recognising and being influenced by social norms. The EMIL architecture (EMIL-A) seems a good candidate for this undertake. EMIL-A has been presented at some length in several papers (for a complete overview, see Conte et al., 2013). In this chapter, we present and discuss an extension of it, i.e. EMIL-I-A, endowed with a set of abilities allowing for different levels of immergence and in particular for *norm internalisation* (see Andrighetto, Villatoro, et al., 2010).

EMIL-I-A is endowed with mechanisms and processes allowing agents to (a) recognise norms; (b) generate new normative representations and according to norm salience act on their grounds; (c) influence other agents by direct communication and by the imposition of different types of punishment[4] and finally (d) internalise norms.

As in any belief–desire and intention (BDI) architecture, EMIL-I-A operates through modules for different sub-tasks (recognition, adoption, decision making and action planning) and acts on mental representations (i.e. goals and beliefs) in a non-rigid sequence. The added value of this normative architecture with respect to existing types, like beliefs–obligations–intentions–desires (BOID) (Broersen, Dastani, Hulstijn, Huang, & van der Torre, 2001) or beliefs–desires–obligations–intentions–norms–goals (BDOING) (Dignum, Kinny, & Sonenberg, 2002), depends on two crucial components, i.e. the *norm recognition* module and the *salience meter*, that, as we show, are necessary elements for allowing norm dynamics to take place in decentralised groups.

The norm recognition module is the main entrance, so to speak, to the architecture. It allows agents to interpret an observed behaviour or a communicated social request as normative and to form the corresponding normative beliefs and goals, thus allowing the norm immergence process to start (for a detailed description of the norm recognition mechanism, see Andrighetto, Campennì, et al., 2010; Campennì et al., 2009).

However, normative beliefs and normative goals are not static representation. Depending on several social or subjective factors, an agent can consider a specific social norm as more or less salient (for himself or herself and for the social group) over time. The more salient is considered, the higher the probability that the norm will be complied with. EMIL-I-A is endowed with a salience control module that allows the agent to understand the relative salience of each norm. The salience module is fed by social and normative information.[5] Those social actions, e.g.

[4] For a description of different types of enforcing mechanisms and their specific effect on people's mind, see Andrighetto and Villatoro (2011), Andrighetto et al. (2013), Giardini, Andrighetto and Conte (2010) and Villatoro et al. (2011).

[5] The resulting salience measure (salience $\in [0-1]$, 0 representing minimum salience and 1 maximum salience) is subjective for each agent, thus providing flexibility and adaptability to the system.

Table 8.1 Norm salience mechanism: cues and weights

Information	Weight
Norm compliance/violation	(+/−)0.99
Observed norm compliance	(+)0.33 × n
Non punished defectors	(−)0.66 × n
Punishment observed/given/received	(+)0.33 × n
Sanction observed/given/received	(+)0.99 × n
Norm invocation listened/received	(+)0.99 × n

behavioural or communicative acts, that are interpreted as compliant with or as defending the norm, make the salience of the norm increase (see Table 8.1).[6]

Conversely, observing unpunished violations makes norm salience decrease, this action being interpreted as a signal that the social group is losing interest in the norm and consequently in its enforcement.

Salience may increase to the point that the norm becomes internalised, i.e. converted into an ordinary goal, or even into an automated conditioned action, a routine. Conversely, it can also decrease below a certain threshold and cease to be a norm for a given agent.

EMIL-A's decisions are influenced by an *aggregation of* utility-based and normative considerations (affected by norm's salience). For example, when deciding whether to comply or not with a norm, on the one hand EMIL-I-As do not want their pay-offs to be reduced by punishment (i.e., utility based considerations), and on the other hand if the norm is perceived as highly salient, their normative motivation will increase and will positively affect the probability of complying with the norm (i.e., norm-based considerations). The decision making of a normative agent is also sensitive to *risk tolerance*: when the perceived punishment probability is below the risk tolerance threshold, the probability of the agent violating the norm decreases.

However beneficial, this process yields (a) *high computational* costs, as each option for action needs to be valued at every time step, and (b) *high social* costs, as norm-abiding agents will behave normatively only in the presence of punishment. Both these costs can be reduced when agents internalise the the norm and start complying with it independent of punishment. As discussed in Andrighetto, Villatoro, et al. (2010), there are several factors favouring norm internalisation, such as consistency, self-enhancing effect, urgency, calculation cost saving and norm salience. In this work, we focus only on the last two conditions. The norm internalisation process takes place when the two following conditions are both satisfied: (1) the norm salience and (2) the cost–benefit calculation for all possible actions *exceeding* a certain threshold.

EMIL-I-As are designed as *parsimonious* calculators: under certain conditions, they internalise norms norms in order to save calculation and execution time. Upholding a norm that has led one to succeed reasonably well in the past—for example, keeping to one's own commitments and therefore being selected as a

[6] These values and their ranking have been extracted from Cialdini et al. (1991), and we are now running laboratory experiments in order to fine-tune them.

partner in transactions—is a way of *economising* on the calculation costs that one would have to sustain whenever facing a new situation.

Once a norm has been internalised, the agent no longer makes the cost–benefit calculation, but observes the norm *automatically*. Nevertheless, the salience mechanism is still active and is continuously updated. In this way, agents can defuse ongoing automatisms and retrieve normative decision making.

What is the value added of EMIL-I-A? As modelled here, norm-internalising agents are expected to outcompete both utility-maximising and simple normative agents.

With regard to the former, EMIL-I-As present an enormous advantage, i.e. observe the norm even when compliance yields a reduced utility for executors. With regard to simple normative agents not endowed with the mechanisms for internalising norms, EMIL-I-As present the advantage of keeping up the rate of cooperation even when punishment is not a sufficient deterrent. In other words, in a population of EMIL-I-As, we expect to observe a high level of compliance and a reduction in the costs for achieving and maintaining it. The higher the salience the likelier the EMIL-I-A agents will adopt the normative goal independent of external sanctions.

At the same time, however, EMIL-I-As are intelligent adaptive agents, and salience is a highly dynamic phenomenon. EMIL-I-A agents are not bound to comply with a norm that is no more in force in their social environment. If salience decreases below a certain threshold, agents will be likely to give it up. But this will not be a sudden effect. Compliance will decrease gradually, and the trend might be inverted easily, keeping the global system performance stable.

Simulating the Effect of Norm Internalisation in Promoting Cooperation

In this section, the performance of the norm-internalising agents, EMIL-I-As, is tested through a simulation experiment that recreates a *social dilemma*. We compare a population of agents endowed with the EMIL-I-A architecture (normative agents able to internalise and de-internalise norms) with two other types of agent: simple normative agents, not able to internalise norms and complying with them only when punishment is a sufficient deterrent, and Instantaneous Utility-Maximiser Agents (IUMAs). IUMAs always choose the action that has given them the maximum benefit in the past.[7] The predictions we aim to test are the following:

- Since IUMAs and simple normative agents follow utility-maximising strategies, they cause cooperation to collapse when punishment rates lower.
- The larger the proportion of EMIL-I-As in the population, the lower the social costs necessary for maintaining a high level of cooperation.

[7] These agents have been implemented ning agents (as in Sen & Airiau, 2007; Villatoro, Sen & Sabater-Mir, 2009).

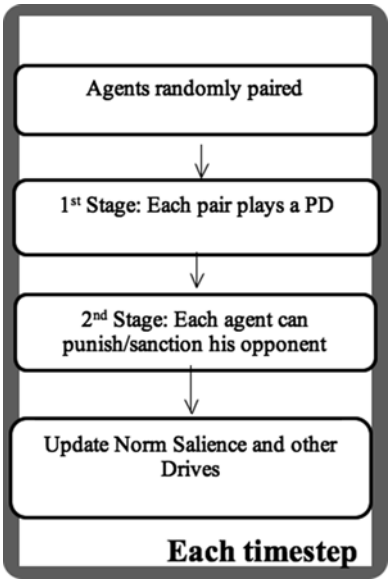

Fig. 8.1 Phases of the game

Table 8.2 Pay-off matrix

	C	*D*
C	3, 3	0, 5
D	5, 0	1, 1

The Game

In this model, agents play a variation of the classic prisoner's dilemma (PD) game, in which an extra stage has been included: after deciding whether to cooperate (*C*) or defect (*D*), agents can also choose whether they want (or not) to punish/sanction the opponents who defected.

Each time step of the simulation is structured in four phases (see Fig. 8.1), which are repeated for a fixed number of time steps. More specifically, these phases consist in the following:

- Partner selection: Agents are paired with other agents randomly chosen from their neighbours.
- *First stage*: Agents play a PD game, with the following pay-offs: $P(C,C)=3, 3$; $P(C,D)=0, 5$; $P(D,C)=5, 0$ and $P(D,D)=1, 1$ (see Table 8.2). It can be exemplified by reference to a situation in which *x* and *y* are competing for a promotion, and each is asked by the employer to give their opinion of the other. If they both say good things about one another (CC), they get the same job part time. If *x* reports positive evaluations on *y* but *y* does badmouthing at the expenses of *x*'s (CD), *y* obtains the position, and *x* gets nothing. If they both spread bad news about each other (DD), they both get an increase in the current salary, but neither gets the promotion. The norm in this scenario is that of abstaining from badmouthing.

- *Second stage*: Agents decide whether to punish/sanction[8] or not the opponents who defected. Punishment works by imposing a cost to the defector, this way affecting its pay-offs and modifying the relative costs and benefits of norm compliance and violation. In addition to imposing a cost for the wrongdoing, sanction also informs the target (and possibly the audience) that the performed action violated a social norm and that conduct is not approved of, thus impacting on both the agent's pay-offs and the process of norm recognition and of salience updating. Only agents who have recognised that a norm of cooperation is in force in their group and that its salience is above a certain threshold use sanction to enforce others' behaviours; otherwise, they resort to punishment.[9]
- *Third stage*: Agents update their utility-based and normative drives.
- Updating normative drive: Both simple normative agents and EMIL-I-As process the normative social information available in their environment. This social information affects directly the norm salience meter, which modifies the normative drive.
- Updating utility-based drive: All the three types of agent calculate the pay-off they received in that round, and this information will influence how they behave in the next round.

Experimental Design

In order to compare the behaviour of simple normative agents, EMIL-I-As and IUMAs, and their relative effect on the achievement and maintenance of cooperation, we have designed a simulation where three different types of agents can interact to perform the same task. All the simulations are populated by a fixed number of agents (=100), with a variable distribution of EMIL-I-As, simple normative agents and IUMAs. From the beginning of the simulation, ten agents are endowed with the cooperation norm, and we refer to them as holders of norms. Since in this work we are not interested in addressing the problem of norm emergence, agents already holding norms from the beginning are necessary to allow the process of norm recognition to start. The results presented in the next section are the average results of 25 simulations.

Experimental Results

The scope of the first experiment is to observe the effect of norm internalisation on norm compliance (in this specific scenario, the cooperation level) even in the absence of punishment. As system designers, we can defuse the punishment/sanction acts inflicted by agents and linearly reduce the probability of their occurrence. Therefore in this first experiment, we vary the proportion of simple normative

[8] The damage of both punishment and sanction is 5 units and the cost is 5/3.

[9] For an analysis of the differences between punishment and sanction, see Andrighetto and Villatoro (2011), Andrighetto et al. (2013), and Villatoro et al. (2011).

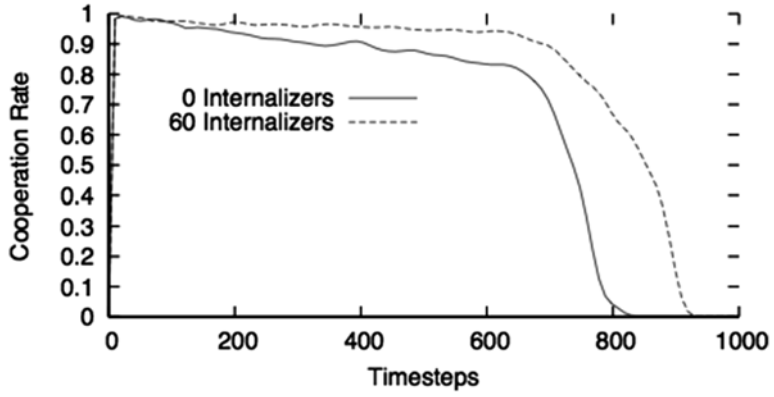

Fig. 8.2 Internaliser dynamic

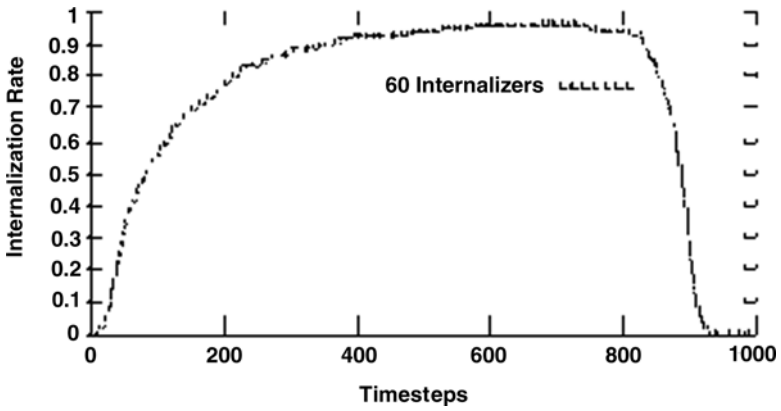

Fig. 8.3 Internalisation rate across time

agents and EMIL-I-As, excluding the existence of IUMAs: one treatment is fully populated by simple normative agents (100) and another one with a majority of EMIL-I-As (60 EMIL-I-As and 40 simple normative agents).

The experimental results shown in Fig. 8.2 prove that the amount of internalisers (EMIL-I-As) is directly proportional to the stability of the cooperation rates: the more the internalisers, the longer the cooperation. The vertical axis shows the average cooperation rate and the horizontal axis the evolution of time in the simulation. The explanation of the phenomenon is found in the dynamics of internalisers: they start behaving as normative agents, and as the punishment probability is above their risk tolerance (see Section "Normative Architecture EMIL-I-A"), they comply with norms. At a certain moment, those that are able to internalise, i.e. EMIL-I-As, do so (as can be seen in Fig. 8.3). However, when the punishment rates decrease, simple normative agents will detect this and start defecting. The more normative agents to defect, the faster the salience (affected by unpunished violations) decreases, also resulting in a faster collapse in cooperation.

Figure 8.3 shows the number of agents that effectively internalised the cooperation norm within the simulation. As discussed previously, during that relative phase, agents behave "automatically," performing no benefit–cost calculation in each interaction. The lack of calculations makes agents who internalised most efficient in terms of execution time.

One important remark about the effects of internalisation concerns the *social cost* to maintain cooperation. In the treatment with zero internalisers, the costs expended for inflicting punishment are much higher (7,369 units per agent) than in the treatment with a majority of internalisers, i.e. 60 agents (3,491 units per agent).

Unlike what might be expected, EMIL-I-As' adaptability is not affected by their automatic performance. The experiment described also shows how internalisers respond to changes in the environment. Once punishment (controlled by the system designer) starts decreasing, internalisers interpret it as a loss of norm salience (see Fig. 8.3). In other words, they consider the lowering of punishment and the consequent increase in the number of unpunished violators as a lack of concern for the social norm. The salience is updated accordingly, driving agents to de-internalise and return after a number of time steps to normative benefit–cost calculation, as is visible in the figure showing cooperation and internalisation rates.

The first experiment shows how internalisers perform in an ideal situation, where all agents (both EMIL-I-As and simple normative agents) have normative beliefs. However, policy makers are interested in less controlled situations, where the interacting agents are much more heterogeneous. The internalisation architecture needs to be tested against other architectures, observing under what conditions cooperation remains stable and at high levels.

In the second experiment, where the probability to punish/sanction decreases as in the previous study, IUMAs interacting with EMIL-I-As are introduced in the same scenario. The population is formed by a constant number of ten EMIL-I-As and with a variable number of IUMAs. In order to keep the population size constant at 100, we introduce as many normative agents as necessary to reach that number of agents (e.g. in the experiment with 30 IUMAs, there will also be the 10 fixed internalisers and 60 other normative agents).

Figure 8.4 shows internalisers adapt to different situations: in situations where no IUMA is present, internalisation rates are much higher than when the number of IUMAs increases. We can see how internalisers can "handle" only a limited number of IUMAs in the society; then, when the number of these is too high (60 IUMAs), internalisers are strongly influenced by them and adapt to the optimal strategy, i.e. they defect, and the cooperation rate collapses.

Moreover, exploiting the maximum of their capacities, EMIL-I-As internalise norms when possible. As can be seen in Fig. 8.5, the internalisation rates achieved when no IUMAs are present is considerably higher than in the rest of the situations. The reason for these phenomena is found in the salience of our internalisers; when IUMAs are present, the number of defections produced is higher, having a negative effect on norm salience and thus preventing internalisers from internalising.

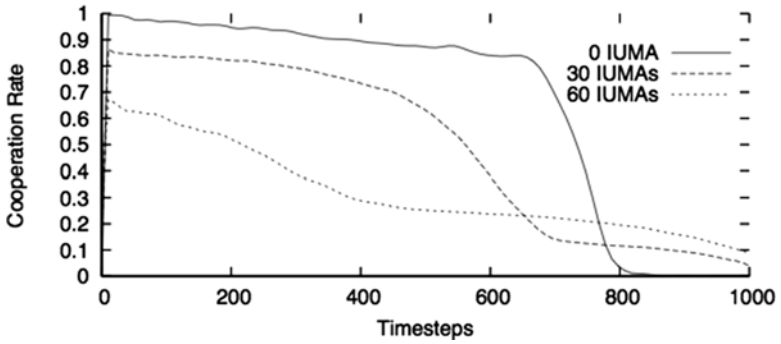

Fig. 8.4 Cooperation rates in mixed populations

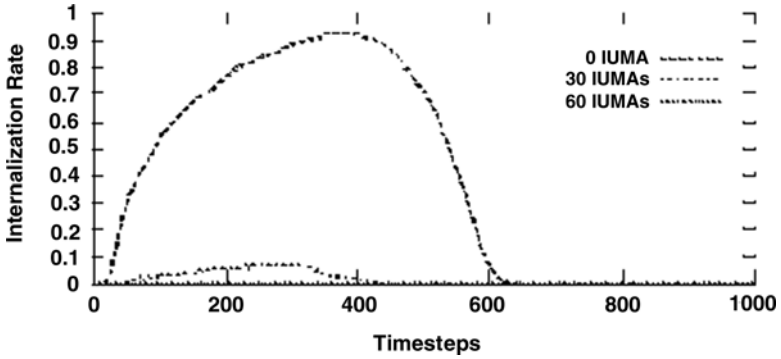

Fig. 8.5 Internalisation rates in mixed populations

Discussion

Our simulation model shows that internalisers are endowed with a rich cognitive architecture allowing them to maintain high cooperation rates even when punishment rates are low. This phenomenon is not observed when dealing with populations with a high number of utility maximisers, leading to a general collapse of cooperation (see Fig. 8.4). We have also observed an interesting phenomenon in populations with a high number of simple normative agents (see Fig. 8.2), which will maintain cooperation only when the punishment rates are above their risk tolerance threshold.

These results would lead us to think that a complete population of internalisers would be the best solution in terms of system performance; unfortunately, this is not true. internalisers do need a certain amount of IUMAs or normative agents to unblock the normative automated actions when necessary. The norm salience mechanism allows the system to maintain a high level of cooperation

even when sanctioning is interrupted. However, when the norm disappears, agents will eventually unblock the automatism generated by the internalisation process and restart the whole process of norm recognition and internalisation. We have also observed that a significant number of internalisers is convenient for the society in general, as they keep the cost of punishment low. Allowing different types of agents to interact provides system policy makers with a tool that can help them predict the dynamics of prosocial behaviour.

Conclusions

In this chapter, we have discussed the necessity for a theory of norm dynamics to account for a dual process, including *emergence* from the bottom up and *immergence* from the top down, illustrating the work done in this direction by the authors.

Essentially, we have argued that norms start to become visible in behaviour once they have shaped and modified the agents' beliefs, goals and expectations and we have provided experimental data confirming this claim. We have represented this process as an iceberg, with the water line standing for the boundary between the observable domain, behaviour, and the unobservable one, the mind. As usual, the portion of the iceberg below the observable level is much larger than its tip, signifying that most of the processing occurs in the mind, below the observation line.

We have modelled norm immergence as a multilevel process and have implemented it on an agent model, EMIL-I-A, where deep levels of norm immergence, namely, internalisation, can be put in action. Simulation results showing how EMIL-I-A works have been discussed, pointing out the individual and social advantages of internalisation in multi-agent applications.

Results obtained so far encourage further developments and applications of our theory of norm internalisation. A research direction deserving further exploration is how to account for the impact of cultural differences on agents' inclination to accept and comply with one or other norm and possibly to internalise it.

Finally, a promising direction of research concerns the internal dynamics of internalisation. In the model presented here, the dynamics of norm salience triggers a special goal dynamics in the mind of the agent: when the norm salience exceeds a given threshold, the normative goal starts to be generated and activated independent of external sanctions. The question of course is why. In Andrighetto, Villatoro, et al. (2010), we hypothesise that anticipatory capacity is a vector of internalisation: agents are likely to experience perturbing feelings and emotions while anticipating external sanctions as consequent to their norm violations. The higher the norm salience, the higher the probability that agents undergo the emotionally unpleasant effects of anticipation, which then start to act as internal sanctions. Future developments of EMIL-I-A might require the design and implementation of a cognitive and emotional model of internal sanctions and of their interaction with the external ones.

References

Andrighetto, G., Campennì, M., Cecconi, F., & Conte, R. (2010). The complex loop of norm emergence: A simulation model. In K. Takadama, C. C. Revilla, & G. Deffuant (Eds.), *Simulating interacting agents and social phenomena: The second world congress, agent-based social systems* (Lecture notes in computer sciences, pp. 17–33). Berlin: Springer.

Andrighetto, G., Campennì, M., Conte, R., & Paolucci, M. (2007). On the immergence of norms: A normative agent architecture. In G. P. Trajkovski & S. G. Collins (Eds.), *Emergent agents and socialities: Social and organizational aspects of intelligence. Papers from the AAAI fall symposium*. Menlo Park, CA: The AAAI Press. Technical Report FS-07-04.

Andrighetto, G., Brandts, J., Conte, R., Sabater, J., Solaz, H., & Villatoro, D. (2013). Punish and voice: Punishment enhances cooperation when combined with Norm-Signalling. *PLoS One, 8*(6), e64941.

Andrighetto, G., & Villatoro, D. (2011). Beyond the carrot and stick approach to enforcement: An agent-based model. In B. Kokinov, A. Karmiloff-Smith, & N. J. Nersessian (Eds.), *European perspectives on cognitive science*. Sofia: New Bulgarian University Press.

Andrighetto, G., Villatoro, D., & Conte, R. (2010). Norm internalization in artificial societies. *AI Communication, 23*(4), 325–339.

Axelrod, R. (1986). An evolutionary approach to norms. *The American Political Science Review, 4*(80), 1095–1111.

Axelrod, R. (1997). *The complexity of cooperation*. Princeton, NJ: Princeton University Press.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. New York, NY: Cambridge University Press.

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theoretical Population Biology, 65*, 17–28.

Boyd, R., & Richerson, P. J. (1988). The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology, 132*, 337–356.

Broersen, J., Dastani, M., Hulstijn, J., Huang, Z., & van der Torre, L. (2001). The BOID architecture. Conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the fifth international conference on autonomous agents, Montreal, Quebec, Canada* (pp. 9–16). New York, NY: ACM.

Campennì, M., Andrighetto, G., Cecconi, F., & Conte, R. (2009). Normal = Normative? The role of intelligent agents in norm innovation. *Mind & Society, 8*(2), 153–172.

Castelfranchi, C. (1998). *Simulating with cognitive agents: The importance of cognitive emergence. Proceedings of the workshop multi-agent systems and agent-based simulation*. Heidelberg: Springer.

Chakrabarti, P., & Basu, J. K. (2010). Emergence of norms in a society of heterogeneous agents influenced by the rules of cellular automata techniques. *International Journal on Computer Science and Engineering (IJCSE), 2*(3), 481–486.

Cialdini, R., Kallgren, C., & Reno, R. (1991). A focus theory of normative conduct. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 201–234). New York, NY: Academic Press.

Conte, R., Andrighetto, G., & Campennì, M. (Eds.). (2014). *Minding norms. Mechanisms and dynamics of social order in agent societies Oxford series on cognitive models and architectures*. Oxford: Oxford University Press.

Conte, R., Andrighetto, G., Campennì, M., & Paolucci, M. (2007). Emergent and immergent effects in complex social systems. In G. P. Trajkovski & S. G. Collins (Eds.), *Emergent agents and socialities: Social and organizational aspects of intelligence. Papers from the AAAI Fall Symposium*. Menlo Park, CA: The AAAI Press. Technical Report FS-07-04.

Conte, R., & Castelfranchi, C. (1995). *Cognitive and social action*. London: University College of London Press.

Conte, R., & Castelfranchi, C. (1999). From conventions to prescriptions. Towards a unified theory of norms. *AI and Law, 7*, 323–340.
Conte, R., & Castelfranchi, C. (2006). The mental path of norms. *Ratio Juris, 19*(4), 501–517.
Conte, R. & Dignum, F. (2001). From social monitoring to normative influence. *JASSS. The Journal of Artificial Societies and Social Simulation, 4*(2), http://jasss.soc.surrey.ac.uk/4/2/7.html.
Dignum, F., Kinny, D., & Sonenberg, L. (2002). From desires, obligations and norms to goals. *Cognitive Science Quarterly, 2*(3/4), 407–430.
Durkheim, E., (1950 [1895]). *The rules of sociological method*. New York, NY: The Free Press.
Epstein, J. M. (2007). Generative social science. Studies in agent-based computational modeling. Princeton, NJ: Princeton University Press.
Faillo, M., Grieco, D., & Zarri, L. (2013). Legitimate Punishment, Feedback, and the Enforcement of Cooperation. *Games and Economic Behavior, 77*(1), 271–283.
Fehr, E., & Fishbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*, 185–190.
Galbiati, R., & Vertova, P. (2008). Obligations and cooperative behaviour in public good games. *Games and Economic Behavior, 64*(1), 146–170.
Giardini, F., Andrighetto, G., & Conte, R. (2010). A cognitive model of punishment. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 1282–1288). Austin, TX: Cognitive Science Society.
Gilbert, M. (1983). Notes on the concept of a social convention. *New Literary History, 14*(2), 225–251.
Gilbert, N. (2002). Varieties of emergence. Paper presented at the Agent 2002. Conference: Social agents: Ecology, exchange, and evolution, Chicago.
Gintis, H. (2003). The hitchhiker's guide to altruism: Gene-culture co-evolution, and the internalization of norms. *Journal of Theoretical Biology, 220*(4), 407–418.
Houser, D., & Xiao, E. (2010). Understanding context effects. *Journal of Economic Behavior and Organization, 73*(1), 58–61.
Masclet, D., Noussair, C., Tucker, S., & Villeval, M. (2003). Monetary and non-monetary punishment in the voluntary contributions mechanism. *American Economic Review, 93*(1), 366–380.
Posner, E. (2000). *Law and social norms*. Cambridge, MA: Harvard University. Press.
Sen, S., & Airiau, S. (2007). Emergence of norms through social learning. In *Proceedings of the international joint conference on artificial intelligence (IJCAI)* (pp. 1507–1512). San Francisco, CA: Morgan Kaufmann.
Ullman-Margalit, E. (1977). *The emergence of norms*. Oxford: Clarendon.
Villatoro, D., Andrighetto, G., Sabater-Mir, J., & Conte, R. (2011). Dynamic Sanctioning for robust and cost-efficient norm compliance. Proceedings of the 22nd international joint conference on artificial intelligence, Barcelona, Spain, July 16–22, 2011.
Villatoro, D., Sen, S., & Sabater-Mir, J. (2009). *Topology and memory effect on convention emergence. Proceedings of the international conference of intelligent agent technology*. Washington, DC: IEEE Press.
Villatoro, D., Andrighetto, G., Brandts, J., Nardin, L. G., Sabater-Mir, J., Conte, R. (In press). The norm-signaling effects of group punishment: Combining agent-based simulation and laboratory experiments. *Social Science Computer Review*, 0894439313511396, first published on December 11, 2013 as doi:10.1177/0894439313511396.
Xenitidou, M., & Elsenbroich, C. (2010). Construct validity and theoretical embeddedness of agent-based models of normative behaviour. *The International Journal of Interdisciplinary Social Sciences, 5*(4), 67–79.
Young, P. (1993). The evolution of conventions. *Econometrica, 61*, 57–84.

Chapter 9
Vulnerability of Social Norms to Incomplete Information

Marco A. Janssen and Elinor Ostrom

Introduction

The ability of groups to self-govern their common pool resources is well documented (Ostrom, 1990). Whether common pool resources are fish stocks or freshwater or forest products, success of self-governance relates to the ability of appropriators to develop trust relationships, monitor and enforce agreements, and communicate among each other.

In this chapter we look at the consequences of a specific element of self-governance, namely, the effect of visibility of the activities on the ability of groups to cooperate. The availability of information about appropriation of actors from a common pool resource can affect the success of self-governance. Resource users may not see each others' actions directly in the appropriation of forests or fish stock. Due to the incompleteness of information resource users need to infer actions of others from the limited information they have.

Incompleteness of information and inference of behavior has been found to be important in other domains of research. A notable example is the misperception of norms related to alcohol use and other drugs (Perkins, 2003). College students, and other population groups, assume that others drink more than they actually do. Hence they expect that the social norm on drinking behavior is a higher use than the actual use. This misperception is caused by limited information. If one sees another student involved in substance abuse, it is assumed to be characteristic of the individual.

Author (E. Ostrom) was deceased at the time of publication.

M.A. Janssen (✉) • E. Ostrom
Center for the Study of Institutional Diversity, Arizona State University, Tempe, AZ 85287-2402, USA

Workshop in Political Theory and Policy Analysis, Indiana University, Bloomington, IN 47408-3895, USA
e-mail: Marco.Janssen@asu.edu

M. Xenitidou and B. Edmonds (eds.), *The Complexity of Social Norms*, Computational Social Sciences, DOI 10.1007/978-3-319-05308-0_9,

Extrapolating behavior of others on occasion to their normal behavior leads to misperceptions. Providing more accurate information on the actual norms of substance use leads to lower levels of substance use (Perkins, 2003).

The frequency in which resource users observe information of others affects the inference they make on their actual use, and this may affect behavior. Before we discuss the natural resource context of our work, it is worthwhile to discuss some definitions of rules and norms as used in our field of political economy. Rules are defined as shared understandings that refer to enforced—by a third party—prescriptions about what actions are *required*, *prohibited*, or *permitted* (Crawford & Ostrom, 1995). Those rules (e.g., law, regulations, contracts) can be defined explicitly on paper or not. In contrast, norms are shared understandings but are not *enforced* prescriptions, meaning that it is not explicitly defined to a third party what to do when a prescription is not met.

Many indigenous systems developed for managing a resource over a very long period of time have found ingenious ways to devise norms and rules that enable them to sustain a resource. There are two big challenges they have to meet. The first challenge is developing a simple set of rules that fit a particular resource system in regard to its boundaries and ecological functioning so as to sustain use over time. Many of the norms or the taboos established by indigenous peoples for controlling their use are "invisible" to outsiders and little understood as to their role in enabling a sustainable resource (Colding & Folke, 2001). The second is creating rules that are easy to follow and easy to determine whether other resource users are also following them or not. When the costs of monitoring performance are kept low, resource users can gain a sense of confidence that rules are being followed on a day-by-day basis without having to invest substantial time and monetary resources in monitoring.

One of the major problems that users of large natural resources face is how to see enough of each other's behavior that they can gain assurance that no one is regularly cheating. When fishers harvest fish from a large territory, farmers withdraw water from a long irrigation canal, or villagers harvest from a large forest, there is no way that they can see what everyone else is currently doing. Many resource management systems developed by local users allocate space and time in a way that authorized harvesters have some assurance that the rules are being followed by others. If everyone is following the rule, then it makes sense for the individual to follow it since the rules ensure that stocks are sustained over time.

In the Maine lobster fishery, for example, rules evolved to allocate permanent spots within a bay to specific fishers (Acheson, 2003; Wilson, Yan, & Wilson, 2007). In this fishery, the map of where everyone is supposed to have the authority to put down lobster pots is common knowledge. If you drew up a pot in your territory that was not yours, this would give you authority to challenge the person who fished there in error. It is said that the first time that a fisher finds a pot illegally placed in their own territory, the fisherman would tie a bow on it to inform the others that they were not following a rule. If that did not work, sanctions could escalate and eventually someone might lose a boat if they did not eventually conform to the rules and norms of their local bay.

Many farmer-managed irrigation systems allocate a particular time to a specific farmer depending on location along the canal and size of farm (and resulting contributions to maintenance of the system) (Burns, 1993; Maass & Anderson, 1986; Shivakoti & Ostrom, 2002). The frequent routine is to allocate a certain time block to a farmer in a specific order either from the head end to the tail or from the tail end to the head of the system (Meinzen-Dick, 2007; Tang, 1992). In either case, that means that when the water transfers from farmer A to farmer B both will be at their distribution point so that farmer A gets as much water as possible but farmer B is able to start watering crops on time. This brings the two most important actors for making this rule enforced to the canal at the same time. Again, this is an ingenious way to enable two key participants to monitor what is happening locally and to enforce any observed rule infractions.

In some of the Alpine commons, quite different norms developed over time related to the harvesting from forested areas. Farmers from the valley that jointly own the Alpine commons work together at a set date to cut an agreed-upon number of trees (Netting, 1981; Stevenson, 1990). Then together they carry the timber and allocate it into approximately equal stacks. The stacks are then randomly assigned to eligible households. Trees cannot be harvested at any other date. This set of norms makes it very easy to control harvesting of stationary resource units from a larger territory. Labor is shared and concentrated on one time a year to cut the trees. It is clear that anyone who cuts at another time is breaking their agreement. Then, everyone has an incentive to make the stacks equal since they will be allocated by lottery to those participating. Again, simple norms that allocate labor and outcomes fairly make it very easy to know when someone is harvesting within or outside their agreements.

The rules developed on most of these indigenous systems also enable participants to chide one another gently if they do find someone who is not following their rules. As mentioned above related to the Maine lobster fishery, these initial gentle chides can escalate over time into graduated punishments that can become pretty severe. Everyone can make an honest error. So there is the problem of gaining assurance that most people are following the rules most of the time as well as giving people a chance to make an error without being thrown out of the community. Most of the long-lasting resource governance systems do involve some form of graduated punishment where the initial reaction is interpersonal discussion about why someone is breaking the rule (Ostrom, 1990). These graduate up to being quite severe punishments, but usually the resource users do not have to impose severe punishments on each other as being called to task in the first place is usually enough to make someone conform. Further, sometimes people just make accidents. Being shown that others notice their accident reassures them that they are in a community that is following the rules in the main and consider the rules to be important. This increases their own trust that cooperating with others and following these rules increase their own long-term benefits.

The examples about suggest that more information leads to better performance. This is not necessarily the case as demonstrated in laboratory experiments. Villena and Zecchetto (2010) show in public good experiments that more specific

information of the actions of other participants in the experiment reduces the level of cooperation. Observing that some individuals do not cooperate can reduce the level of cooperation in the group. Similarly, Janssen (2013) shows in a spatially explicit common resource experiment that more information leads to a more rapid decline of the resource. In both studies there is no communication. Due to the lack of communication and the limited information, initial optimistic expectations of the behavior of others might persist. With more detailed information, participants are able to identify others who are less willing to cooperate, leading to a decline of cooperation.

In this chapter we represent information availability by changing the vision of agents in the spatially explicit resource. A larger vision means that agents have information about a larger share of the resource and about more other agents. We will discuss this in more detail when we discuss the model.

The important thing about these self-organized systems is that the rules they design and adapt over time fit the ecology and social conditions in which they exist and they have worked effectively for long periods of time. Why do they work? Well, they are a reliable and low-cost mechanism for allocating resource units on systems that extend over space. If one did not have rules like this, one would need to hire monitors to regularly patrol the area in order to get an overview of what was going on across the entire resource. This can be a very expensive effort as well as one that involves conflict and challenge because the only person who has the relevant information about what is going on around the resource is the guard. Individual resource users can only challenge whether the guard is correct or not. In the systems where users have developed norms that enable them to rotate harvesting activities across time and space or involve easy-to-identify spatial allocations, the resource users themselves can monitor each other and assess whether others are following the rules or not.

Methods

The research on collective action and the commons is interdisciplinary and multi-method (Poteete, Janssen, & Ostrom, 2010). In the early days of the field rational choice theory and the use of mathematical models dominated the field. This approach led to a convincing argument that people are not able to overcome the tragedy of the commons (Hardin, 1968). Privatization, taxes, or other interventions were needed to avoid overharvesting. However, since the 1980s, scholars from disciplines like anthropology, political science, history, psychology, sociology, and economics are involved in unraveling the puzzles of the field using multiple methods. Case studies are used to show that people are not always trapped in a tragedy of the commons. In fact, they are often successful to self-govern. Moreover, comparative analysis shows the importance of monitoring and enforcement over the specific property rights implemented (Dietz, Ostrom, & Stern, 2003).

Controlled experiments have been used to test hypotheses on human behavior, while agent-based models are used to test develop alternative theories informed by case studies and experiments (Poteete et al., 2010). This is also the approach used in this chapter.

With agent-based models we refer to computational representations of autonomous agents who interact with each other at a microlevel leading to broader level patterns. Agent-based models are information-processing algorithms based on various assumptions about the cognitive ability of the individual agents and the topology of their interactions. Agent-based models have been used since the early 1980s to study collective action problems (e.g., Axelrod, 1984).

The simulation model we present in this chapter assesses in a simplistic way the trade-offs between the cost of information and the effect on resource management. We only focus on social norms of the agents and do not include monitoring and enforcement. We will start the simulation with a social norm shared by all agents that will lead to the cooperative solution. Then we will investigate whether the social norms will remain followed when individuals will receive mixed signals due to incomplete information.

The model is based on an experimental environment we use to study how communication and punishment (together or separately) affect the harvesting rates of users (Janssen, Holahan, Lee, & Ostrom, 2010). In these experiments we see participants develop shared norms on when and where to harvest resources. We basically explore the consequences of the vision radius of agents on the actions of the agents. In future studies we will extend this modest model with explicit monitoring and enforcement actions to study the fit of effective monitoring arrangements for different types of ecologies.

Model

We start with a resource system of $N \times N$ cells. A cell can contain one resource unit that can be appropriated by one of the M agents (Fig. 9.1). The resource renewal rate is density dependent. The probability that a resource unit will reappear on an empty cell increases as the number of adjacent cells with tokens increases. The probability p_t is linearly related to the number of neighboring cells with tokens $p_t = p \times n_t/8$ where n_t is the number of neighboring cells containing a green token, 8 is the number of neighboring cells, and p is 0.01.

The optimal strategy to maximize earnings in the longer term as a group is to harvest only resource units when there are four or more resource units in the eight neighboring cells (Janssen et al., 2010). This can be explained since the model is a spatial explicit version of the traditional logistic growth model of renewable resources that has the maximum growth rate at 50 % of the carrying capacity. The replenishment rate of the whole resource is highest if every empty cell is surrounded by four tokens.

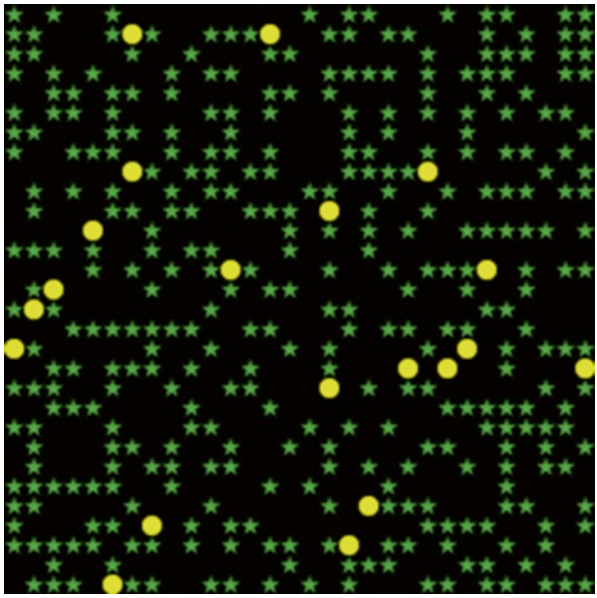

Fig. 9.1 Screenshot of the model environment with $N=30$ and $M=20$

Agents make one step each time step. Agents can only move left, right, up, and down. Thus the list of possible actions is {left, right, up, down}. Only one agent can be at a cell at one moment. Hence when an agent is located next to a cell the set of possible actions is reduced. For each direction a score is calculated of the number of resource units within the vision of the agent in that direction. Vision is defined as the radius around the agents for which resource units as well as other agents are visible. The cells nearby are linearly weighted more than the cells faraway. When an agent has a vision 5 a resource unit in the next cell is weighted five times a token five cells away. The agent will move to the direction with the highest score. With a small probability ($p_r=0.1$) an agent remains moving forward and does not reconsider direction.

When an agent is located on a cell with a resource unit it considers to harvest the resource unit. An agent counts the number of tokens on the eight surrounding cells and uses a threshold T_A when to harvest the token. If T_A is equal to four, the agent will only harvest a token if there are four or more tokens on the eight neighboring cells. If all agents will follow this rule there will be no overharvesting of the resource.

In our simulations all agents start with a threshold equal to 4. This represents the condition that all agents share the same social norm that will lead to the cooperative optimum solution. They may adjust the harvesting threshold norm if they observe resource scarcity. Resource scarcity is defined as the density lower than the density expected equal to the threshold value divided by 8. Hence with a threshold value of four agents expect a density of 50 % tokens.

If agents experience a lower threshold than expected they reduce their threshold T in the following way:

$$T_t = \lambda_1 \times T_{t-1} + (1-\lambda_1) \times (d \times 8) \tag{9.1}$$

where d is the density of resource units within the radius of vision, and λ_1 is the weight of threshold T_{t-1} when relative resource scarcity is observed. If λ_1 is equal to 1, the threshold will not change.

The threshold T_A which is actually used by the agent in the decision making is an integer since the number of tokens on the eight neighboring cells is an integer. The T_t value is rounded to T_A. Hence when T_t drops from 4 to 3.6, the agent still uses a threshold T_A equal to 4.

Agents can also observe that the density is in line with their expectations. This happens when the density is higher or equal to (T_A/8). The threshold recovers back to the value of T_A when it observed agents behaving as expected:

$$T_t = \lambda_2 \times T_{t-1} + (1-\lambda_2) \times T_A \tag{9.2}$$

where λ_2 is the weight of threshold T_{t-1} when relative resource scarcity is not observed. Confirmation bias (Nickerson, 1998) leads people to weigh confirmation to expectations more than they weigh surprises. To include this into our model we assume that $\lambda_1 \geq \lambda_2$. This means that a surprise leads to a smaller adjustment than a confirmation. If we include skeptical agents $\lambda_2 \geq \lambda_1$.

There is a second way agents can adjust their threshold. We assume that agents can make an error, defined as harvesting a token when the number of tokens of neighboring cells is below T_A of that agent. If an agent i observes agent j taking a token below the threshold T_A agent i uses, agent i will adjust the threshold level T_t in the following way:

$$T_{t,i} = \eta_1 \times T_{t-1,i} + (1-\eta_1) \times T_{\text{used},j} \quad \text{when } T_{\text{used},j} < T_{t-1,i} \tag{9.3}$$

where $T_{\text{used},j}$ is the number of tokens around the eight cells agent j just collected. The parameter η_1 is the weight of threshold $T_{t-1,i}$ when an agent j is observed who uses a lower threshold to harvest resource units than agent i. If parameter η_1 is equal to 1, agents will not adjust their threshold if they see somebody harvesting resource units at lower densities then they approve:

$$T_{t,i} = \eta_2 \times T_{t-1,i} + (1-\eta_2) \times T_{A,i} \quad \text{when } T_{\text{used},j} \geq T_{t-1,i} \tag{9.4}$$

Similar to λ_1 and λ_2 we assume that $\eta_1 \geq \eta_2$.

In our simulations we will explore the consequences of the occurrence of a selfish agent in the group. A selfish is defined as the agent who starts with a threshold value equal to 0. A documented version of our implemented model can be found at http://www.openabm.org/model/2284/version/1/view (Table 9.1).

Table 9.1 Parameter values used for the model

Parameter	Description	Value
p_r	Probability of random direction	0.1
V	Vision	[1, 10]
λ_1	Weight of threshold when relative resource scarcity is observed	[0.9, 1]
λ_2	Weight of threshold when relative resource scarcity is not observed	$2\lambda_1 - 1$
η_1	Weight of threshold when an agent is observed with a lower threshold	[0.9, 1]
η_2	Weight of threshold when an agent is observed with the same or larger threshold	$2\eta_1 - 1$
p_e	Probability that a harvesting agent uses a lower threshold	[0, 0.02]

Results

A number of simulation runs are performed with a 30×30 resource and 20 agents. The first set of simulations explores the relationship between vision and resource size when agents do not make errors. Resource size is looked at since it is an indicator for the level of cooperation in the population. Further, we investigate the impact of one cheater in a population of conditional cooperators. The resource size is measured as the average of the last 100 time steps of a 1,000 time step simulation. For each combination of (λ_1, η_1) we ran 100 simulations for both none and one selfish agent in the population. A selfish agent is defined as an agent who harvests a token whenever a token is available for harvesting.

Figure 9.2 shows that when λ_1 and η_1 are equal to 1, and there is no selfish agent, an increase in vision leads to a slight reduction of the long-term resources. Although the agents will not change their thresholds they use for harvesting (Fig. 9.3), a larger vision will enable them to find more resources. This level of the long-term resource size is the baseline to compare the effects of threshold changes.

When there is one egoist in the group, while λ_1 and η_1 are still equal to 1, there is a significant reduction of the resource size (line (1,1,1) in Fig. 9.2). The effect is the largest for vision equal to 3, which is caused by the movement decisions of the agents. With a large vision, agents will not be moving into areas with the selfish agent so that it can recover.

If η_1 is equal to 0.9 and λ_1 is equal to 1, the agents will reduce the threshold due to observing other agents using thresholds lower than their own threshold. Without an egoist the line in Figs. 9.2 and 9.3 is the same as the case of (1,1,0). When there is an egoist (1,0.9,1), there is a significant reduction of the level of resources. With smaller vision, agents are less likely to see other agents and selfish agents will be able to escape the attention of other agents (Fig. 9.3). With vision larger than three cells, the selfish agents will be observed, and this leads to a decline of the average threshold value (Fig. 9.3). With large vision agents will also derive confirmation that selfish behavior is rare, leading to a modest increase of the threshold.

When only λ_1 is 0.9 and η_1 is equal to 1, we see that the resource size is small with a small vision and increasingly large with vision. In this scenario the threshold is only adjusted when scarcity is observed. With a small vision it is more common to

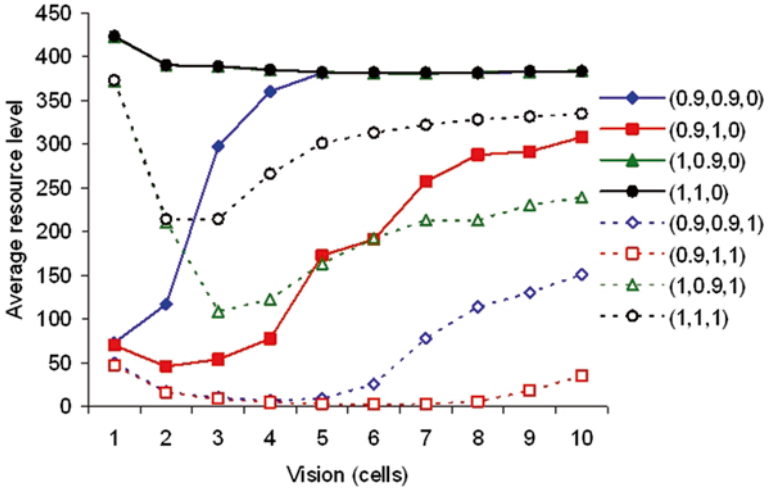

Fig. 9.2 The average resource size of the last 100 time steps of 1,000 time step simulation for each condition of (λ_1, η_1, number of egoists) for ten different levels of vision

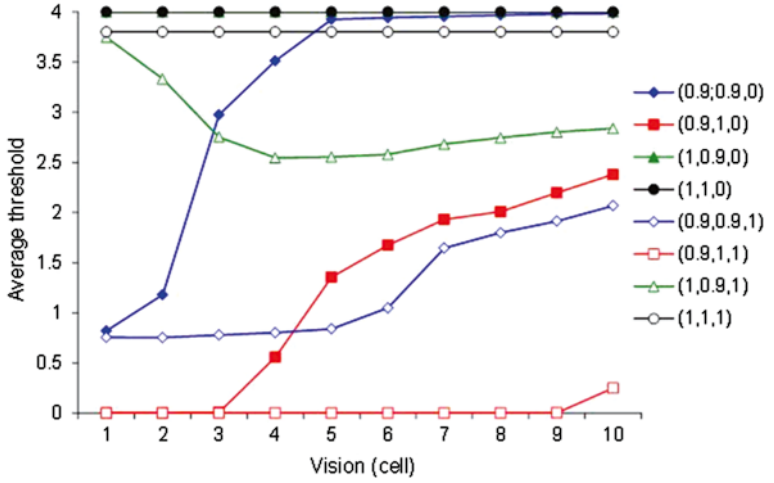

Fig. 9.3 The average threshold of the last 100 time steps of 1,000 time step simulation for each condition of (λ_1, η_1, number of egoists) for ten different levels of vision

have an underestimation of the resource availability. Since a decline of the resource threshold is faster than the recovery, a higher vision increases recovery and the value of the resource size. When there is a selfish agent, the level of resource size is low for all levels of vision. The resource is overharvested since agents experience local overharvesting, and the selfish agent reinforces this inference.

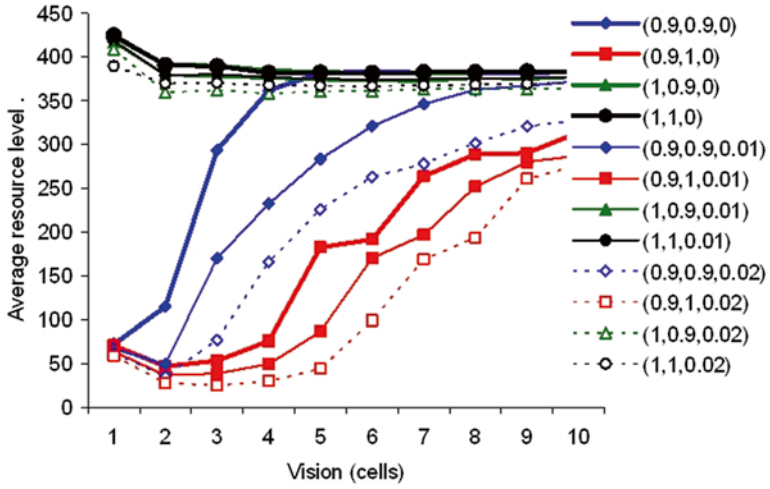

Fig. 9.4 The average resource size of the last 100 time steps of 1,000 time step simulation for each condition of (λ_1, η_1, p_e) for ten different levels of vision

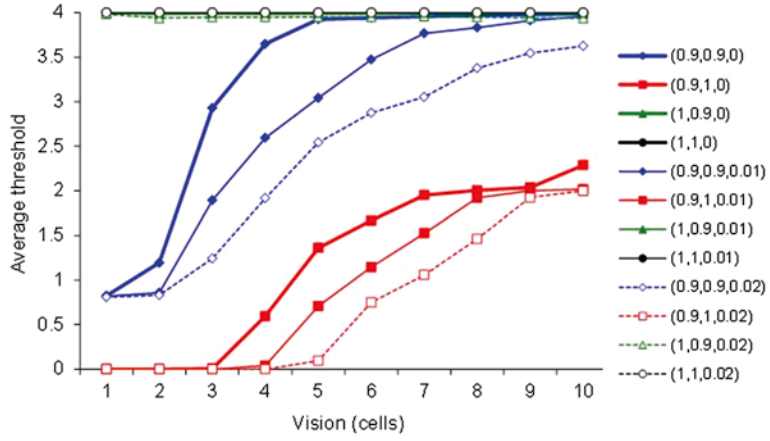

Fig. 9.5 The average threshold level of the last 100 time steps of 1,000 time step simulation for each condition of (λ_1, η_1, p_e) for ten different levels of vision

When agents adjust their thresholds for both observation of scarcity and cheaters ($\lambda_1 = 0.9$ and $\eta_1 = 0.9$) the resource decline is smaller compared to adjusting for resource scarcity only. The reason is that the frequency to observe a cheater is relatively small, and thus agents increase their thresholds when they observe no cheating behavior.

Subsequently we perform a number of simulations without cheating agents, but with a rate of selfish behavior, p_e is equal to 0, 0.01, and 0.02 (Figs. 9.4 and 9.5).

We see that when agents do not update their thresholds due to local resource scarcity ($\lambda_1 = 1$), the errors made by cooperative agents have a minor impact. We can see this in Figs. 9.4 and 9.5 since all lines (1,*,*) have a threshold around 4 and a resource level around 400.

On the other hand, when agents do reduce their threshold when they observe more resource scarcity than expected ($\lambda_1 = 0.9$), we see an increased level of resource availability for larger vision. Furthermore, the sensitivity to errors is much larger. When p_e is 0.02 the threshold is lower than for $p_e = 0.01$ or $p_e = 0$.

Conclusions

Ecological context can affect the amount of information agents derive in order to make decisions to follow the norm in a group. The size of the system and the change in the elevation within a system both decrease the information that agents can obtain about resource conditions and the harvesting activities of other users. Using an agent-based model of a spatially explicit renewable resource we observe that the increased ability to observe the state of the resource increases the capacity of a group of conditional cooperators to maintain the level of cooperation. However, increased ability to observe other agents cheating will reduce the level of cooperation if the other agents make errors or purposely try to cheat.

Future laboratory experiments with a spatially explicit resource based on Janssen et al. (2010) will enable us to improve assumptions on how people adjust their behavior to observations of relative resource scarcity and selfish behavior. Janssen (2013) find that without communication, limited information leads to less overharvesting, while with communication, limited information leads to lower performance and more overharvesting.

In our research program over the years we have found it extremely productive to use multiple methods including field studies, simulations, and experimental research to study a common question from multiple perspectives (Poteete et al., 2010). Future research will combine agent-based models and experiments in the lab and the field to understand what aspects of information on the actions of others are sensitive to the performance of the group.

The model did not include explicit monitoring and enforcement actions. In the examples discussed in section “Introduction” information availability affects the type of rules crafted by the resource users. In future work we will study what type of harvesting rules can evolve that fit the ecological context. For example, when vision is limited, an effective strategy for some resource, such as a forest, might be to harvest together. This is not the optimal strategy when vision is unlimited and thus agents may spread the harvesting pressure more evenly.

This exercise illustrates the sensitivity of cooperation to the level of information that agents derive. Due to limited information conditional cooperative agents infer assumptions about what others do and this affects their own behavior. When less information is needed to monitor the resource, norms on when and where to harvest

are less critical. This illustrates that for effective institutions one need to take into account the costs of monitoring. With higher costs of deriving information, one needs to develop rules that make the mutual monitoring more effective.

Behavioral research has shown that humans often follow norms of conditional cooperation in social dilemmas. Information about the behavior of others may indicate that a "bad apple" exists in the community that can spoil the whole bunch or that cooperative behavior is confirmed. The effect of information is likely dependent on the existing norms within the population. If participants think that the norm is to conform a modest harvesting level, more information may lead to a decline of the norm towards overharvesting if overharvesting is observed. On the other hand, if one is agreeing to reduce harvesting, more information that confirms that others obey the norm may reinforce this norm. More research—experimental and theoretical—is needed to tease out these counterforces.

Acknowledgements This work is supported by a grant from the National Science Foundation (SES-0748632). The authors thank two anonymous reviewers for helpful comments.

References

Acheson, J. M. (2003). *Capturing the commons: Devising institutions to manage the Maine Lobster Industry*. Hanover, NH: University Press of New England.

Axelrod, R. (1984). *The evolution of cooperation*. New York, NY: Basic Books.

Burns, R. D. (1993). Irrigated rice culture in monsoon Asia: The search for an effective water control technology. *World Development, 21*(5), 771–789.

Colding, J., & Folke, C. (2001). Social taboos: "Invisible" systems of local resource management and biological conservation. *Ecological Applications, 11*(2), 584–600.

Crawford, S. E. S., & Ostrom, E. (1995). A grammar of institutions. *American Political Science Review, 89*, 582–600.

Dietz, T., Ostrom, E., & Stern, P. (2003). The struggle to govern the commons. *Science, 302*, 1907–1912.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*, 1243–1248.

Janssen, M. A. (2013). The role of information in governing the commons: Experimental results. *Ecology and Society, 18*(4), 4.

Janssen, M. A., Holahan, R., Lee, A., & Ostrom, E. (2010). Lab experiments for the study of social-ecological systems. *Science, 328*, 613–617.

Maass, A. & Anderson, R. L. (1986). … and the Desert shall rejoice: Conflict, growth and justice in arid environments. Malabar, FL: R. E. Krieger.

Meinzen-Dick, R. (2007). Beyond panaceas in water institutions. *Proceedings of the National Academy of Sciences, 104*(39), 15200–15205.

Netting, R. M. C. (1981). *Balancing on an Alp*. Cambridge: Cambridge University Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.

Ostrom, E. (1990). *Governing the commons. The evolution of institutions for collective action*. Cambridge: Cambridge University Press.

Perkins, H. W. (2003). The emergence and evolution of the social norms approach to substance abuse. In H. W. Perkins (Ed.), *The social norms approach to preventing school and college age substance abuse: A handbook for educators, counselors, and clinicians* (pp. 3–17). San Francisco, CA: Jossey-Bass.

Poteete, A., Janssen, M. A., & Ostrom, E. (2010). *Working together: Collective action, the commons, and multiple methods in practice*. Princeton, NJ: Princeton University Press.

Shivakoti, G., & Ostrom, E. (2002). *Improving irrigation governance and management in Nepal*. Oakland, CA: ICS Press.

Stevenson, G. G. (1990). *The Swiss grazing commons: The economics of pen access, private, and common property*. Cambridge: Cambridge University Press.

Tang, S. Y. (1992). *Institutions and collective action. Self-governance in irrigation*. Oakland, CA: ICS Press.

Villena, M. G., & Zecchetto, F. (2010). Subject-specific performance information can worsen the tragedy of the commons: Experimental evidence. *Journal of Economic Psychology, 32*, 330–347.

Wilson, J., Yan, L., & Wilson, C. (2007). The precursors of governance in the Maine Lobster Fishery. *Proceedings of the National Academy of Sciences, 104*(39), 15212–15217.

Part III
Evaluating Complex Approaches to Norms

Chapter 10
The "Reign of Mystery": Have We Missed Something Crucial in Our Experimental and Computational Work on Social Norms?

Flaminio Squazzoni

This book is needed for three reasons. Firstly, it helps us to realise the importance of understanding dynamic social norms by studying the interplay of micro and macro aspects. Secondly, it allows us to appreciate this difficult challenge and the need for advanced experimental and computational approaches. Thirdly, it is here at the right moment.

Let us first look at the present state of our society. Now, individualisation is on the ebb, people are fragmented into social groups that develop, overlap and disband even across virtual spaces and large-scale social and technological changes are dramatically modifying the material and cultural bases of our lives. New institutional and normative equilibria will probably soon take place at various levels, e.g. society, the economy and politics. In this situation, understanding how social norms emerge, under what conditions they persist or change and how we could promote or inhibit them is essential to ensure that groups and communities can regulate themselves for the collective good.

Previous studies have suggested that we follow social norms for a variety of reasons. We do so in view of certain material or reputational benefit, as others expect, as it is generally good, as we learnt to do so by relatives and peers, unconsciously and by habit or simply to save time for more rewarding, pressing or emotional activities. Game theory, sociology, social psychology, cognitive sciences and economics have explored all the various angles of social norms (e.g. Dubois, 2002; Gintis, Bowles, Boyd, & Fehr, 2005).

Certain authors have indicated that social norms could be more fruitfully understood by formalised models, which leads us to simplify, abstract and experiment (e.g. Corten & Buskens, 2010; Ostrom & Walker, 2003). Others have defended the idea that social norms and their dynamics could be better understood by description

F. Squazzoni (✉)
Department of Economics and Management, University of Brescia, Brescia, Italy
e-mail: flaminio.squazzoni@eco.unibs.it

M. Xenitidou and B. Edmonds (eds.), *The Complexity of Social Norms*, Computational Social Sciences, DOI 10.1007/978-3-319-05308-0_10,

and history (e.g. North, 2005). I agree with this book that formalised and quantitative research, including computational and experimental research, is crucial to disentangle the social mechanisms of norms and to integrate theory and observation better. Although this does not disqualify the importance of qualitative and historical descriptions of norm emergence and change, I believe that significant explanatory advances can be made especially by evidence-based formalised theory (e.g. Squazzoni, 2008; Timmermans, de Haan, & Squazzoni, 2008).

Having said this, my contribution aims to discuss two main arguments. First, I would like to reconsider the social component of the book's equation: "cognition + social context = complexity of social norms." My idea is that this component has been an invisible guest in most book contributions. In the first part of this chapter, I focus on the role of social structures in influencing social norms and I provide experimental and simulation findings that indicate that norms are sensitive to "who interacts with whom." My understanding is that the idea of "social context" is too vague if not anchored to concrete social interaction structures.

Secondly, I would like to discuss the coherence of research strategies followed by this book's contributions and their expected results. I agree with Andrighetto et al. that experimental research can look only at the "observational" side of normative facts, not at their mental counterparts. This is evident also in the brilliant experimental chapters by Welsey Perkins on misperceptions, Cristina Bicchieri and Hugo Mercier on deliberation, Christine Horne on norm enforcement and Marco A. Janssen and Elinor Ostrom on commons. On the other hand, unlike Andrighetto et al., I have serious doubts about looking at unobservable, mental processes from a hard cognitive approach or that it is sufficient to understand social norms. Certain recent advances in neuro-economics and neurobiology have shown that individuals' normative behaviour is less cognitive and more emotional and social (see also Elster, 2007, 2009). I would also like to question whether agent-based modelling (ABM) is the most appropriate strategy to look at these mental facts.

Before continuing, I would also like to confess that I am a sociologist. As such, I am interested in explaining large-scale social outcomes from agent interaction in social structures. In my work, the behavioural and cognitive components of individual behaviour are instrumental to explain social outcomes and not an end in itself, nor a tribute to the truth (Coleman, 1990). This must be said as other colleagues might have different objectives and study social norms from other angles. However, my contribution to this book is to provoke a constructive debate and so I will be less panegyric than the book may otherwise really deserve.

The "Social Component" or the Invisible Guest

Most contributions here have emphasised the importance of embedding social norms into a social context. The editors have indicated that one of their main purposes is to reinvigorate the study of norms by looking at time and context. I suppose that all contributors would agree in saying that cognition and individual behaviour

is intrinsically social and that norms strongly depend on social interaction and are influenced by the social context. However, my impression is that this "social component" has been poorly elaborated. Therefore, I would like to suggest that materialising this "social component" into concrete and structured social interactions could improve our understanding of social norms.

Numerous sociological studies have shown that the mechanics of social interaction, for which social structure is largely responsible, dramatically influence the emergence of social norms. This is because rational motives and normative foundations of individual behaviour are not clearly separable and so are context dependent (e.g. Bowles, 2008; Gintis et al., 2005). Schelling (1978), Powell (1990), Coleman (2000) and Burt (2005) among others showed that social structures influence social norms because they embody certain mechanisms of social interdependence. Indeed, the mechanics of social contacts play a crucial role in determining and generalising social norms as individuals are extremely sensitive towards the behaviour and opinion of other individuals whom they are in contact with (e.g. Centola & Macy, 2007; Granovetter, 1978; Watts, 1999).

For instance, it is widely acknowledged that a dense, stable and relatively closed social structure, where individuals tend to interact frequently and repeatedly, tends to reduce free riding and favour norm convergence and persistence. This type of structure provides room for reputation-building strategies, magnifies behavioural signals, permits reciprocal behaviour's monitoring and makes relatively low-cost social sanctions possible (e.g. Buskens, 2002). This may happen in criminal gangs but also in neighbourhood associations, workplaces or among groups of friends. Obviously, the situation drastically changes if we look at more open and flexible social structures, such as markets. In this case, social norms are insufficient to ensure cooperation and powerful institutional arrangements such as contracts are needed to help us to reduce transaction costs and share the cost of social control (e.g. Cook, Hardin, & Levi, 2005).

Recent studies have found that even a network's configuration matters for norm emergence. For instance, Buskens, Corten, and Weesie (2008) built a simulation model that showed that network density influences the way behaviour develops: the higher the density, the stronger the influence of the initial behavioural distribution on a behaviour's emergent distribution. Moreover, they found that if social networks are initially segmented, as usually happens in reality for socio-historical reasons, the coexistence of different norms and even their polarisation is more probable.

By combining experiment and computational work on coordination games in dynamic networks, Corten and Buskens (2010) showed that any norm equilibrium is extremely sensitive to social influence. This is because individuals are strongly influenced by whoever they are in contact with and use the observed behaviour of neighbours to predict (often erroneously) the behaviour of unknown partners (e.g. Salganik & Watts, 2009). They also found that less efficient norms tend to persist when the social structure consists of dense networks as conformity prevails. On the other hand, more efficient norms can emerge if social structures endogenously develop with individual choices, i.e. if networks are flexible and breaking/creating links is economically and informationally possible.

We found a similar finding in an experimentally grounded model, where we replicated the experimental behaviour of subjects in an agent-based model so as to look at the macro implications of micro-scale behaviour. We wanted to understand the dynamic interplay of social interaction and social structure in trust situations (Bravo, Squazzoni, & Boero, 2012). First, 108 subjects played a repeated investment game, where subjects were randomly coupled to play as investors or trustees. Investors were asked to decide how much of their endowment to send to trustees, who in turn received the amount tripled by the experimenters and had to decide how much to return to the former. After the experiment, we estimated the behaviour of each subject in each round through a statistical model that looked both at the individual trust propensity of subjects and their capability of reacting upon experience.

Subsequently, we used these experimental data to calibrate an agent-based model that reproduced the experiment. First, we tested the impact of various network structures on cooperation. Then, we introduced the possibility that agents broke and created links according to a simple happiness threshold function. The investors/trustees were happy when trustees/investors returned/invested more or the same as in the previous interaction, and when happy they continued to interact with the same partners.

While static network configurations did not significantly alter the experimental results, dynamic networks based on partner selection significantly improved cooperation and fairness. This was due to the fact that, while norm abiders benefitted from more interactions and links and were more profitable by ensuring in turn higher profitability for their partners, the "bad apples" were isolated over time. In short, the social structure dynamically adapted to positive outcomes of social interaction, which in turn strengthened the functional configuration achieved over time by increasing the contact density between "good guys." This confirmed the fact that certain features of the social structure can play a soft social control function that helps individuals to defend positive norms and self-regulate their interaction for collective benefit.

The influence of social structures is also important in that it can magnify certain behavioural propensities that individuals show even in less complex network structures. In Boero, Bravo, Castellani, and Squazzoni (2009), we investigated the impact of reputation on trust and cooperation in structures based on random encounters. Starting from a typical investment game, such as above, we added the possibility that investors could rate trustees' behaviour at the benefit of subsequent investors, who knew the rating of the trustee with whom they were matched before the investment. Ratings were expressed in terms of "positive," "neutral" and "negative" trustee. Obviously, these investors' ratings were subjective, as investors' opinion on trustees depended on their own expectations and their level of investments. This meant that reputational information incorporated imprecise, even misleading information on trustees' real intentions and therefore had to be cautiously considered by investors. However, the simple fact that information on subject behaviour at time t_0 was available at time t_1 dramatically influenced both investors' and trustees' behaviour by ensuring more reciprocal trust. This was for various reasons.

In agreement with previous studies (e.g. Keser, 2003), we found that trustees returned significantly more when they were under rating as they were rationally motivated by reputation building. Although reputation was formulated under potential bias, negative emotions and subjectivity of investors, it was rational for trustees to achieve a good standing in view of future benefits from investors' decisions. On the other hand, we also tested a treatment where trustees' reputation was available to investors only after their decision, so that reputation building for trustees was ruled out. Even in this case, trustees behaved more fairly than when there was no reputation.

Moreover, we tested a treatment where even the investors were under reputation by trustees. In this case, trustees received the amount of resources sent by investors with whom they were matched as well as their past reputation. This information should not have any consequences on subjects' behaviour, as trustees should be influenced by the amount sent by investors and not by investors' previous reputation. Also in this case, there was no room for reputational building strategies by investors.

Our results showed that in all conditions, adding reputational information created more cooperation, irrespectively of its consequence for individual material pay-offs. Our conclusion was that subjects were influenced by being under judgment more than any induced, more rational, stimulus–response incentive. Once introduced and irrespective of its concrete economic value for the interaction, reputation implied that subjects framed the game as a moral problem and played more fairly (Kahneman & Tversky, 2000). It is worth noting that this occurred even if interaction was anonymous and communication was forbidden.

This was also confirmed by studies that looked at the role of gossip for cooperation (e.g. Dunbar, 1996, 2004). An experimental study showed that subjects were influenced by gossip even when they were also able to use other sources of information, including direct observation (Sommerfeld, Krambeck, Semmann, & Milinski, 2007). Piazza and Bering (2008) experimentally tested a modified version of the dictator game, where subjects were asked to distribute an endowment between themselves and an anonymous second party. Half of the participants were told that the second party would be discussing their decision with a third party. Their results showed that individuals dramatically overreacted to the possibility of being the subject of gossip by increasing their fairness even if the negative consequences of gossip were hardly predictable. Again, subjects were influenced by a mix of strategic reasoning and attention towards their own social approval.

The same interplay of rational motives and normative foundations was found in certain studies where subjects' mental processes were monitored through functional magnetic resonance imaging. For instance, Hsu, Anen, and Quartz (2008) showed that different brain regions activated whenever subjects faced a difficult trade-off between rational considerations based on efficiency motives and widespread social norms such as equity. More specifically, they found that a specific brain region (the *putamen*) responds to efficiency, while a second one (the *insula*) responds to equity. A third region (the *caudate/septal subgenual*) encodes a unified measure of these two motives and is probably linked with the resolution of the trade-off. Moreover, they found that a behavioural measure of individual differences in

inequity aversion correlates with the activity measured in the equity encoding regions. We can also imagine similar psychological mechanisms acting on reputation and leading to rational reputation-building actions that are, at least partially, separated from more social cognition-driven behaviours.

Recent experimental work has shown that subtle reputation-related cues significantly modified individuals' behaviour. As in our case, those cues were especially linked with the possibility of being observed. For instance, Haley and Fessler (2005) showed that the presence of stylized eyespots on computer desktops used for the experimental sessions significantly increased the generosity of players in a dictator game despite no differences in actual anonymity. In another work, conducted in a real-world setting, Bateson, Nettle, and Roberts (2006) found a similar effect of apparently unimportant cues of being watched. Their results showed that people put nearly three times as much money in an "honesty box," used to collect money for drinks in a university coffee room, when the cost of the drinks was displayed on a board along with a picture of eyes staring at the consumer than when the notice included a flower control picture.

It is curious to note that the effect of being watched is so striking that subjects even reacted when the "observer" was not human. The participants in another experiment contributed significantly more to a public good when a robot picture, which obviously represented a machine but endowed with two large eyes, was placed on their computer desktops (Burnham & Hare, 2007). This is to say that individuals in typical social situations would react more emotionally than cognitively and rationally.

In Bravo, Squazzoni, and Takács (2012), we extended these experimental designs to include intermediaries, who were asked to observe the exchange between investors and trustees and rate trustees' behaviour for investors. By doing so, we added a further layer of complexity as we transformed the typical dyadic trust relationship between investors and trustees in a triadic interaction.

As George Simmel argued in his famous piece on the significance of numbers for social life (Simmel, 1950), this extension has serious consequences. By adding a third element to a dyadic relationship, various processes can take place that were previously impossible, such as positive or negative intermediation, impartial opinions and more moderate passions (see also Coleman, 1990). For this reason, it is important to understand which incentives and social norms ensure cooperation between three actors in different roles. To do so, we tested various incentive schemes, by aligning intermediaries' pay-offs to investors or trustees and by excluding any material incentive. We also tested the same incentive schemes by keeping the role of structures fixed.

First, we found that the presence of intermediaries increased cooperation compared with dyadic reputation-based interaction as trustees were more trust responsive when rated by an intermediary. We also found that individuals were more sensitive to fairness and equity of the exchange when material incentives of intermediaries were ruled out. The triadic interaction structure, if combined with role alternation, provided room for indirect reciprocity motives that increased fairness and equity. This meant that, for intermediaries, being helpful to investors by

keeping the evaluation standard of trustees' fairness high at time t_0 was essential to expect reliable reputational information on trustees by investors playing as intermediaries at time t_1.

Furthermore, the lack of material incentives ensured intermediaries' neutrality as intermediaries' opinion was seen by the other figures as more credible as disinterested. It was essential however that the interaction structure included role rotation, as this ensured reciprocity strategies and allowed subjects to understand the implications of their decision in each role better.

To conclude this brief excursus, we can say that experimental and simulation research on social norms had two main findings. Firstly, understanding the strength of social norms without considering the social structure effect can bring partial conclusions, as the structure is a carrier of social influence and social influence is very important for norm emergence and persistence.

Secondly, most experimental work on social interaction confirmed that even in a cold social context such as the lab, individuals are influenced by moral sentiments and emotions, such as indignation, shame, envy and gratitude related to social approval (e.g. Elster, 2009; Gintis et al., 2005). This can explain the important role of reactive behaviour even in an interaction context, such as a strategic game, where experimenters intentionally induce the self-interest of subjects and the rational component of subjects' behaviour should dominate.

It is worth noting that neural investigation also confirms this point. Recent studies have shown that individuals' normative behaviour in social interaction can be understood more in terms of simple rewards and emotional schemes than as the result of complex psychological or cognitive factors (e.g. Glimcher, Colin, Russell, & Fehr, 2008). For instance, De Quervain et al. (2004) examined the neural basis for altruistic punishment of defectors in an investment game similar to our own one. The only modification was that investors had the chance of punishing unfair trustees by bearing a certain cost. They found that subjects derived personal satisfaction from punishing norm violators by activating the brain area related to rewards. They also found that individual differences in the motivation for this altruistic behaviour (e.g. normative vs. hedonistic motives), which are overemphasised in psychology, were irrelevant.

The social consequence of this behaviour was more important in that it created evolutionary advantages for good guys and preserved the well-being of the group by reducing opportunities for unfair behaviour. This means that moral emotions are crucial to understand the strength of social norms and that something "precognitive" could even take place as individual behaviour seems more influenced by biological and social factors (see also Rilling et al., 2002). This could also account for the importance of reactive behaviour that we all (wrongly) surprisingly see in the lab.

This discussion also has important implications for the next point of my chapter, as it brings us to consider the advantages of tighter integration between experimental and computational approaches and field and bio-neural work (e.g. Harmon-Jones & Winkielman, 2008; Ross, Sharp, Vuchinich, & Spurret, 2008). This means that we need to question whether ABM research and traditional experimental behavioural approach in the social sciences are sufficient in looking at the puzzle of social norms.

The Limits of ABM Research on Social Norms

Some chapters, such as Goldspink's, Elsenbroich's and Andrighetto et al.'s, have discussed the limitations of social simulation when looking at social norms, especially from a cognitive point of view. I would now like to outline that there has also been excessive emphasis on the capability of social simulation to look at all the subtle angles and implications of social norms. After reviewing about 15 years of ABM research in the social sciences (Squazzoni, 2010), I have realised that the most influential and widely acknowledged ABM applications have looked at the macro-level impact of agent interaction starting from simple micro mechanisms (Squazzoni, 2012). This could be for two reasons: it may reflect the fact that understanding the "bottom-up emergence" of social norms is one of the founding constituencies of all social sciences, or it may be given that ABM research is especially suitable to do so. I believe that both interpretations are true.

Social scientists have been fascinated by ABMs as they allow us to observe the large-scale, macro-level behaviour of systems based on agent interaction. It must be remembered that this has always been one of the most important challenges for any social science right from its conception (e.g. Coleman, 1990; Schelling, 1978). Before the advent of ABMs, we lacked methods and research technology to do so. This is the secret of the success of ABMs in the social sciences (Epstein, 2006; Epstein & Axtell, 1996; Hedström, 2005), not the fact that they could help us to understand the mental aspects of individual behaviour or to look at sophisticated cognitive processes behind individual behaviour.

ABM research cannot look at the whole picture of social norms including cognition, without losing its key feature, which is the study of macro-consequences of agent interaction. Indeed, we must consider that there is a trade-off between complexificating the cognitive components of models of social outcomes and understanding the impact of agent interaction on social outcomes. First, the extent to which the sophistication of cognitive component of a model should be pushed is a pragmatic choice and not an ontological starting point (Gilbert, 2005). This means that adding sophisticated cognitive properties to agent interaction models is useful only when it has been proven that more simplified and general assumptions are insufficient to explain the social outcome of interest. Secondly, we must consider that any sophistication comes at the cost of explanatory capacity, transparency and replicability of models, with dramatic consequences for cumulativeness and scientific advancement (Squazzoni, 2012).

Another criticism against over-sophisticated cognitive models is empirical validation. It must be recognised that it is difficult to produce testable findings on complex socio-cognitive aspects of social interaction at the level needed to look at social norms. So, one of the main challenges for all cognitive-sided contributions here is to understand how their findings could be empirically tested by observation and how to do so by remaining within the boundaries of experimental and computational research. I believe that neuroscientific research, and even more traditional qualitative research, could be especially suitable to look at inner cognitive mechanisms of individual behaviour and to validate cognitive explanations. I would like

to see computational and experimental research become more integrated with these approaches.

All in all, my impression is that computational and experimental approaches have severe limitations when looking at social norms. First, through experimental research and simulation, we can only observe the link between interaction and typically simplified agent behaviour and the possible consequences of this for large-scale agent interaction systems. This has also been outlined by Cristina Bicchieri and Hugo Mercier, and Marco A. Janssen and Elinor Ostrom, in this book: without field experiment and empirical research it is difficult to understand the micro mechanisms that cause individual behaviour. This implies that experimental research is mostly used to observe deviations from pre-constituted theories, such as rational choice predictions, rather than to find generative, causal explanations.

The reason for this is because it is impossible to look at emotions, unconscious reactions, effects of prior exposure and socialisation on individuals in the lab. This implies that in order to understand social norms, we often need to call for something outside the lab, not fully covered by the experimental design or indirectly understandable only "par différence." This gives us an idea of certain limitations of experimental research when dealing with mental and cognitive causes of individual behaviour, while again, this type of research is decisive for looking at social interaction on a small scale and in a simplified "stimulus–response" framework.

It is worth pointing out that these limitations do not disqualify experimental and computational approaches to social norms. I do not want to be misunderstood. In my view, these approaches are absolutely necessary. The problem is that they are not yet sufficient to understand the entirety of social norms. This requires cross-fertilisation of various methods, something that Poteete et al. recently called "working together" (2010). In this example, various methods, including field experiments, ethnographic research and surveys, were integrated with lab experiments and ABM to understand an important issue, i.e. collective action and commons. This cross-methodological work was inspired by a common framework and pursued a common explanatory goal. In this way, research could overcome the gap of observation scales that penalises its development and understand the link between local knowledge of social interaction and global implications better (Squazzoni, 2012). To conclude, this "working together" is a good example of the type of research that we should try to do more often in the field of social norms.

Conclusions, Obviously Partially Inconclusive

Here, I have tried to discuss how to embed the perspective of this book sociologically and I have pleaded for better integration between various types of research. I have suggested looking more carefully at the role of social structures in influencing social norms, so as to give a more concrete dimension to the idea of the "social context." I have also insisted on certain limits of strong cognitive approaches and suggested the importance of looking at more simple micro mechanisms of individual

behaviour. Examples include recent studies in neurobiology and neuroeconomics, whose findings could be positively integrated into experimental behavioural and ABM research. Finally, I have also outlined certain problems of current experimental work on social norms.

To conclude, the answer to most of these "critical arguments" will possibly come from research technology development. In the future, we may be capable of integrating empirical work, neural investigation, with experimental and ABM research, which are now pursued in parallel, by observing, for instance, human beings under magnetic resonance interacting in large-scale systems or by having access to a large amount of data on human behaviour at low cost and in real time. Maybe in the future, the quality of data available for social science research will significantly improve and the new social media will allow us to amplify our recourse to experimental research, so that theory and observation will become more integrated. As always happens in the history of science, innovation is strongly dependent on technology progress. Let us hope so.

But for the time being, we can say that the challenging issues presented in this book already demonstrate that disciplinary and research method barriers should be viewed as the result of institutional, organisational and historical processes of the science system rather than something that truly reflects important epistemological reasons. Certain examples in this book have allowed us to envisage future developments, but let us be more courageous and try to accelerate the pace of this "working together" attitude from today.

References

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters, 2*, 412–414.

Boero, R., Bravo, G., Castellani, M., & Squazzoni, F. (2009). Reputational cues in repeated trust games. *Journal of Socio-Economics, 38*, 871–877.

Bowles, S. (2008). Policies designed for self-interest citizens may undermine 'The Moral Sentiments': Evidence from economic experiments. *Science, 320*, 1605–1609.

Bravo, G., Squazzoni, F., & Boero, R. (2012). Trust and partner selection in social networks: An experimentally grounded model. *Social Networks, 34*(4), 481–492.

Bravo, G., Squazzoni, F., & Takács, K. (2012). Intermediaries in trust: An experimental study on incentives and norms. Collegio Carlo Alberto Notebooks, No. 240: http://www.carloalberto.org/assets/working-papers/no.240.pdf

Burnham, T. C., & Hare, B. (2007). Engineering human cooperation: Does involuntary neural activation increase public goods contributions? *Human Nature, 18*, 88–108.

Burt, R. S. (2005). *Brokerage and closure: An introduction to social capital*. New York, NY: Oxford University Press.

Buskens, V. (2002). *Social networks and trust*. Dordrecht: Kluwer Academic Publishers.

Buskens, V., Corten, R., & Weesie, J. (2008). Consent or conflict: Coevolution of coordination and networks. *Journal of Peace Research, 45*, 205–222.

Centola, M., & Macy, W. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology, 113*, 702–734.

Coleman, J. (1990). *Foundations of social theory*. Cambridge, MA: The Belknap Press of Harvard University Press.
Coleman, J. (2000). Social capital in the creation of human capital. In E. L. Lesser (Ed.), *Knowledge and social capital: Foundations and applications*. Boston, MA: Butterworth Heinemann.
Cook, K. S., Hardin, R., & Levi, M. (2005). *Cooperation without trust?* New York, NY: Rusell Sage Publications.
Corten, R., & Buskens, V. (2010). Co-evolution of conventions and networks: An experimental study. *Social Networks, 32*, 4–15.
De Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science, 305*, 1254–1258.
Dubois, N. (2002). *Sociocognitive approach to social norms*. London: Routledge.
Dunbar, R. I. M. (1996). *Grooming, gossip and the evolution of language*. Cambridge, MA: Harvard University Press.
Dunbar, R. I. M. (2004). Gossip in evolutionary perspective. *Review of General Psychology, 8*, 100–110.
Elster, J. (2007). *Explaining social behavior*. Cambridge, MA: Cambridge University Press.
Elster, J. (2009). Emotions. In P. Hedström & P. Bearman (Eds.), *The Oxford handbook of analytical sociology* (pp. 51–71). Oxford: Oxford University Press.
Epstein, J. M. (2006). *Generative social science. Studies in agent-based computational modeling*. Princeton, NJ: Princeton University Press.
Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies. Social science from the bottom up*. Cambridge, MA: The MIT Press.
Gilbert, N. (2005). When does social simulation need cognitive models? In R. Sun (Ed.), *Cognition and multi-agent interaction: From cognitive modeling to social simulation* (pp. 428–432). Cambridge: MA: Cambridge University Press.
Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (Eds.). (2005). *Moral sentiments and material interests: The foundations of cooperation in economic life*. Cambridge, MA: The MIT Press.
Glimcher, P. W., Colin, C., Russell, P., & Fehr, E. (Eds.). (2008). *Neuroeconomics: Decision making and the brain*. London: Academic.
Granovetter, M. S. (1978). Threshold models of collective behavior. *American Journal of Sociology, 83*, 1420–1443.
Haley, K. J., & Fessler, D. M. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior, 26*, 245–256.
Harmon-Jones, E., & Winkielman, P. (Eds.). (2008). *Social neuroscience: Integrating biological and psychological explanations of social behavior*. London: Guilford Press.
Hedström, P. (2005). *Dissecting the social. On the principles of analytical sociology*. Cambridge, MA: Cambridge University Press.
Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science, 320*, 1092–1095.
Kahneman, D., & Tversky, A. (2000). *Choices, values, and frames*. Cambridge, MA: Cambridge University Press.
Keser, C. (2003). Experimental games for the design of reputation management systems. *IBM Systems Journal, 42*, 498–506.
North, D. (2005). *Understanding the process of economic change*. Princeton, NJ: Princeton University Press.
Ostrom, E., & Walker, J. (Eds.). (2003). *Trust and reciprocity: Interdisciplinary lessons from experimental research*. New York, NY: Russell Sage.
Piazza, J., & Bering, J. M. (2008). Concerns about reputation via gossip promote generous allocations in an economic game. *Evolution and Human Behavior, 29*, 172–178.
Poteete, A. R., Janssen, M. A., & Ostrom, E. (Eds.). (2010). *Working together: Collective action, the commons, and multiple methods in practice*. Princeton, NJ: Princeton University Press.
Powell, W. W. (1990). Neither market nor hierarchy: Network forms of organization. *Research on Organizational Behavior, 12*, 295–336.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron, 35*, 395–405.

Ross, D., Sharp, C., Vuchinich, R. E., & Spurret, D. (2008). *Midbrain mutiny: The picoeconomics and neuroeconomics of disordered gambling*. London: The MIT Press.

Salganik, M. J., & Watts, D. J. (2009). Social influence. The puzzling nature of success in cultural markets. In P. Hedström & P. Bearman (Eds.), *The Oxford handbook of analytical sociology* (pp. 315–341). Oxford: Oxford University Press.

Schelling, T. (1978). *Micromotives and macrobehavior*. New York, NY: W. W. Norton.

Simmel, G. (1950). *The sociology of Georg Simmel*. Glencoe, IL: The Free Press.

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., & Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *PNAS, 104*(44), 17435–17440.

Squazzoni, F. (2008). A (computational) social science perspective to societal transitions. *Computational and Mathematical Organization Theory, 4*, 266–282.

Squazzoni, F. (2010). The impact of agent-based models in the social sciences after 15 years of incursions. *History of Economic Ideas, 2*, 197–233.

Squazzoni, F. (2012). *Agent-based computational sociology*. Chichester, West Sussex: Wiley.

Timmermans, J., de Haan, H., & Squazzoni, F. (2008). Computational and mathematical approaches to societal transitions. *Computational and Mathematical Organization Theory, 4*, 391–414.

Watts, D. J. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology, 105*, 493–527.

Chapter 11
Three Barriers to Understanding Norms: Levels, Dynamics and Context

Bruce Edmonds

Understanding the phenomena that we summarise as "social norms" is difficult, and the reasons for the difficulty go to the heart of the social sciences. It may turn out that some social phenomena might be understandable independently of the particularities of human cognition that in some sense some patterns of social behaviour would be common to different species or even societies of artificial entities. It is possible that some social phenomena can be modelled using a clever, but ultimately simple and general, model that captures some aspect of behaviour across different situations and cases. It might even be that some social phenomena emerge in a purely "upward" manner from the complex interaction of individuals.[1] But social norms are not in these categories. Rather, social norms are rightly called complex. Indeed, they are not complex in only one sense, but in multiple ways. They are highly dynamic in that social norms arise, hold sway to differing extents and in different ways for a whilst and fall into disuse. They act in parallel, where different (and even conflicting) social norms may be pertinent at the same time, each relating to different social groups and origins. They arise and derive their stability from both an upwards emergent process and a downwards "immergent" process, with societal level constraints upon the individuals. They are highly context dependent, with different norms impacting upon behaviour in very different ways in different situations, in some almost impossible to go against and others strictly optional. They involve deep and particular features of human cognition. They seem to have evolved (both biologically and socially) as a result of human sociality as well as enabling its development. Their actions and impacts involve conscious reasoning and perception as well as unconscious processing. They are not rational in any simplistic sense but nor are they irrational—rather they lie at the intersection of our individual and

[1] I think cases where any of these turn out to be the case will be rare, at best. However one cannot rule them out.

B. Edmonds (✉)
Centre for Policy Modelling, Manchester Metropolitan University, Manchester, UK
e-mail: bruce@edmonds.name

M. Xenitidou and B. Edmonds (eds.), *The Complexity of Social Norms*, Computational Social Sciences, DOI 10.1007/978-3-319-05308-0_11,

societal purposes and adaptions. They touch upon our individual and collective identities, both in playing a part in their formation as well as resulting from these. Many of the chapters in this book point out some of these complexities and start to explore them.

Given all these complexities, the question naturally arises whether the idea of a social norm is a useful one. It is becoming increasingly apparent that what may seem simple, a social norm, is far from this. However, people do seem to be able to identify social norms and talk about them without apparent difficulty, especially when they are violated. Furthermore, different people agree to a surprising extent on what the social norms that pertain to a situation are and are able to use their knowledge to manipulate others (e.g. get acceptance to a proposition when it would be rude to refuse). In other words, they seem to have an intersubjective social *reality* that belies the academic difficulties in their precise identification. Thus we are faced with an apparent paradox—what people find so easy to identify, navigate, understand and talk about seems impossibly complex to those who study them. Of course, this paradox is not unique to social norms, but I would argue that some of the reasons we (as academics) find social norms difficult are rooted in how we approach trying to understand social phenomena. This does not mean that these are the only barriers to understanding them, but they are substantial. The three academic habits I will focus on here are static analyses, single-level understanding and context-independent models. To express these as positives, I am arguing that what is needed are dynamic analyses, inter-level understanding and context-sensitive modelling. These suggestions, as ways forwards, are shared with several of the chapters in this book at different times. These are each discussed in turn.

Inter-level Understanding

In all fields of life, people learn specific skills and approaches and then seek to apply them. Sometimes these imply a particular perspective on whatever is being studied, since it is natural to use the cognitive tools that one has better access to. Someone who tries to understand social phenomena using aggregate data and statistical models will probably make better progress at increasing understanding from the aggregate level, looking for (sometimes weak) clues as to what factors might be influencing other factors, in general—general trend connecting some specific issues at the macro-level. An ethnographer who records accounts of how individuals behave and cope with their situation will have a very different view—a rich, semantic, micro-level, specific view. A cognitive scientist will have yet another view, trying to understand how the thought processes within an individual combine to result in individual behaviour in very specific cases—an extremely micro perspective.

An unfortunate tendency of academic fields is that they seek to "insulate" themselves against having to attend to the results of other fields they may not fully

understand or be comfortable with.[2] Thus an ethnographer may ignore the results of aggregate statistics because it does not come to grips with the full complexity of individual behaviour within the appropriate environment, whilst the statistical modeller may find an ethnographic account deficient because it is "only" a specific example, which might be insufficient to inform policy. Both may ignore the results from psychology or cognitive science, simply because it is so alien to them, requiring a lot of background knowledge to even read. That different traditions study social norms from very different perspectives is understandable; one cannot be master at all trades. However these traditions do not collaborate with each other very much, developing such different styles, aims and languages that even a cross-field conversation is difficult to hold. That human society, and in particular human social norms, would be independent of the details of human cognition seems *prima facie* implausible; however, it is largely overlooked in social science.

However, it seems increasingly clear that understanding what social norms are, how they arise, how they impact on observed behaviour and how they fall into disuse *requires* an inter-level approach: not just looking at social norms from different perspectives but how the different levels interact to produce social norms. Several chapters in this book emphasise this.

Giulia Andrighetto et al. (Chap. 8) argue that a dual-cognitive-social account is essential to the very nature of norms. They produce and specify a cognitive model that allows for norms to be guessed at (fallibly) from observing others' behaviours which may become a normative belief. Only when the collective behaviour and normative beliefs coincide is a social norm effective.

The most basic, inter-level factor that is considered is that our beliefs about others and their norms might not accord with what is in fact the case (Perkins, Chap. 2). That is, our beliefs about what is normal in terms of behaviour do not accurately reflect the behaviour of others that our beliefs concern. Here the connection between basic cognition and the society the cognitive agents compose is important to the dynamics of norms—in particular, how they first get established.

A similar issue is highlighted in Marco Janssen's and Elinor Ostrom's contribution (Chap. 9). Here the transparency and completeness of information are shown to be necessary for maintaining norms that enhance sustainable use of commons. In Chap. 3, Cristina Bicchieri and Hugo Mercier look at the relationship of beliefs and the empirical and normative expectations, using this as the fulcrum for interventions as to how to facilitate a change in a person's behavior and beliefs. This builds upon Bicchieri's book (2005) that elaborates the cognitive mechanisms and social conditions necessary for normative beliefs to arise.

Corinna Elsenbroich (Chap. 5) crosses the social and cognitive levels explicitly in her argument for the importance of "we-intentionality" in human normative behaviour. This is different from individuals with their own intentionality that manage to coordinate their action, since those concerned attribute the goals and intentionality to the collective to which they belong.

[2] I only have anecdotal and personal evidence for this, but it is something I have repeatedly observed across several different fields.

However Goldspink (Chap. 4) goes deepest into the many levels that may come into the production of social norms. His analysis of the emergence of norms involves many abilities and processes that exist at a variety of levels, from the biological origin of what is meaningful and "good," through affect and emotion, agency and identity up to cultural tools and language. He sketches some of the characteristics of a simulation that might capture some of what is necessary in such an emergentist account.

Although none of these multi-level accounts are complete, one can detect substantial areas in which they agree, pointing towards a dual-cognitive-social account of norms that holds out the possibility of accounting for many of the normative phenomena that we observe.

Dynamic Analyses

Social norms are often presented (e.g. to children) as a simple social fact. Added to this is the longevity of some social norms, where factors seem to conspire to "lock in" the norm long after the conditions that lead to its formation have ceased to hold. Thus the social norm to "eat everything on your plate" in the United Kingdom outlasted the WWII austerity years by several decades. For this reason, and others, some accounts of norms have been more concerned with defining what they *are* rather than how they come about and change—focussing on their ontology rather than their ontogeny. However, a good definition of a phenomenon is one that is in concert with how that phenomenon comes about. Thus, a comparative feature definition of a species was okay as a guide as to the appropriate clustering of species, but it was only with an understanding of how different species evolve, including their genetics, that a deeper definition was arrived at. It might be that some norms are so stable and persistent that their dynamics do not need attending to, but even then we would not be able to say *why* it is so. A dynamic analysis would be able to account for why and when it makes sense to identify a complex of social phenomena as a norm. Thus ontogeny is important even if one is only concerned with ontology.

The existence of "life cycle" of a social norm suggests several questions: How do the complexes we recognise as social norms come about? What happens when the social norms that are relevant to different groups and/or identities are inconsistent with each other? How and under what circumstances can social norms change or be reinterpreted? Why do some social norms fall into disuse and become disregarded? Although far from giving a complete answer to any of these, many of the chapters are moving towards partial answers.

In Chap. 3, Cristina Bicchieri and Hugo Mercier look at what might be levers of change and how norms might be shifted via discussion and deliberation. They argue that "individuals will abandon a shared social norm only if they believe that others are changing, too." This means that people have to change norms together in

a coordinated manner. When deliberation of a norm (against more deeply held beliefs) would lead others to change, but it is disadvantageous to do so if others are perceived to disagree, then a process of slow diffusion, very slow at first, may reach a tipping point where it very quickly disappears.

Brigitte Burgemeestre et al. (Chap. 7) look at how norms might change via what they call "bounded norm emergence." They identify two dynamic processes: a bottom-up one of norm emergence through a developing consensus and a top-down one of producing and checking that specific versions of rules suitable in their context are consistent with the current institutional view. They dissect some very specific case studies to examine how these processes are operating there.

This simultaneous two-way dynamic of bottom-up emergence with top-down immergence is further developed and simulated in Chap. 8, by Giulia Andrighetto et al. Here two simulations to explore the resulting dynamics are discussed, both of which, in their different ways, compare the situation where agents do and do not have the cognitive requisites to support social norms. This enables them to demonstrate that a dual-cognitive-social account of norms is necessary but also then allows some of the dynamics, for example what happens in mixed populations, to be explored.

Another simulation study is discussed in Chap. 9 of Marco A. Janssen and Elinor Ostrom. In particular when and how norms may collapse depending on the amount of information available to participants are examined. The effect of increasing or decreasing information spread is not simple—an increased ability to observe the state of the resources increases cooperation, but an increased ability to observe others apparently cheating may decrease it.

Context Dependency

Which social norms seem to effectively constrain behaviour seems to vary sharply with the situation that pertains. What is acceptable during war is different from what is acceptable at a shared meal. Norms also seem to be relative to different groups; for example, people may have one norm derived from their professional identity as a lawyer and a conflicting norm to overlook trespassing done by fellow protesters about a planned road. One person may never question a norm (say, against public nudity), whilst another feel that this is an oppression of personal freedom that needs to be publically contravened. Indeed one of the paradoxes of social norms is that, on the one hand, their efficacy is context dependent, whilst on the other they seem to be important in producing a certain kind of regularity in behaviour across societies.

It is fair to say that context "haunts" the social sciences. Whilst it is obvious that much human behaviour—*especially* human social behaviour—is highly context dependent, as individual researchers we need to generalise across particular

situations if our research is to be useful to anybody else. In particular, the context dependency of social phenomena has been seen as inimical to being a *science*, when this is conceived of as *only* being concerned with general laws. Certainly if one wishes to use current analytic or statistical mathematical models then it is hard to include much about how things differ in each context.

Several chapters in this volume point out and face this problem as it touches on normative behaviour. Wesley Perkins (Chap. 2) does not assume that behaviour is context independent, but rather targets his research at particular groups and issues, looking for the in-context response (as far as this is practical using survey methods). Of course, even within his target there will be people making different decisions based on the specific personal sub-context that they inhabit. The statistical approach he uses to detect significance within his target does not take into account these personal sub-contexts, but rather simply looks for general statistical facts concerning the results of his surveys, etc. However, the research he describes does help us to understand how the cognitive context of the individual is constructed, in particular how the set of assumptions about the "background" normative assumptions are adopted. His chapter vividly shows that the contrast between what generally holds in terms of behaviour and what individuals assume to be the norm can significantly differ, driving the dynamics of the norms across the group (by either reinforcing an existing norm or establishing a new norm).

In Chap. 3, Cristina Bicchieri and Hugo Mercier address a similar theme to Perkins, looking at how change in belief and general behaviour needs to be coordinated if change is to be effected by interventions or persuasion. In her highly influential[3] book on norms (Bicchieri, 2005) she looks at much greater detail about how normative behaviour must be context dependent, with different norms being triggered in a context-sensitive manner.

An important part of the relevant context for people is, of course, the social context. Christine Horne (Chap. 6) highlights the crucial interdependence of group members in determining what norms are considered relevant and, in particular, which of these will be enforced. The key contextual information of which social grouping (or associated identity) is relevant to an individual at any particular time seems both crucial to understanding norms as well as relatively neglected in their study.

Brigitte Burgemeestre et al. (Chap. 7) address context head on, looking at how it helps to drive norm development. They show how norms can be adapted to particular circumstances as the result of seeking to apply a norm in those circumstances. At the same time the set of specific versions of the norms influence the direction of the more general norm. Although this chapter looks at a specific kind of situation (regulations concerning distance driven by lease car drivers in the Netherlands) similar kinds of normative development are observed across the world, including international human rights, norms for behaviour in international business meetings and norms that govern scientific behaviour.

[3] Well we would say that, wouldn't we!

New Affordances and Methodological Mixes

The computational techniques that pervade complexity science allow for some new[4] approaches to the three problems outlined above. It should be emphasised that I do not think that these are panaceas, but rather (as I will discuss further below) techniques that can be used alongside existing techniques.

After the problem of taking into account what may be happening at various levels (within people, within groups, within society, etc.[5]) there is the problem as to how the processes at these different levels interact with those at other levels. This is sometimes characterised as the problem of relating the micro- and macro-levels (although it could clearly also be the meso-macro or micro-meso problems, etc. Sawyer, 2005). Keeping track of the relationships both within and between levels quickly becomes extremely complicated and beyond the ability of the human mind to track. Whilst there are some established techniques that touch on such multi-level phenomena (e.g. multi-level statistical models) they do not allow for the inter-level processes to be included, just the co-existence of the levels. Agent-based modelling, in contrast, explicitly addresses the relationship between micro- and macro-levels. This includes the emergence of effects from the micro- to the macro-level (where the word "emergence" implies that this is precisely something that cannot be predicted without such a model) but also the "immergence" or the downward causation of the macro-level back to the individual. Although in physics it is almost universal to only allow the upward direction of emergence, there is no reason for simulation models in the social scientists to so restrict themselves (Conte, Edmonds, Moss, & Sawyer, 2001).

Whilst there are many established techniques for studying and understanding dynamic systems and processes, this is hard if these are complex, for example, where there is emergence as discussed immediately above. Of course, simulation modelling and dynamical formal approaches can be of help here, but mostly this seems to be more a matter of attitude. It is simply easier to restrict oneself to a snapshot of what is happening or look for structures that have some permanence about them. Here it is more the influence of complexity science which has taken to understanding and visualising dynamic processes that seems to be more pertinent. A dynamic view of processes is becoming the norm!

However computational power and storage are certainly making dynamics studies easier, facilitating the logging of large amounts of social data tagged both in time and space (e.g. Birkin et al., 2010). The advent of the computer means that new ways of studying dynamic social phenomena are becoming feasible.

---

[4] Whilst many of the ideas in complexity science have been around for some time, including in the social sciences, the advent of accessible computer power has enabled the development of new tools and approaches that start to make applying these ideas feasible.

[5] Whilst the idea of "levels" is clearly a simplification, this terminology is sufficient for the purposes of this discussion.

Finally, computational techniques can also help with dealing with context dependency. Since the raw data can be preserved, manipulated and distributed digitally there is no need to unnecessarily or prematurely generalise away contextuality in order to communicate about it or model it, since computational techniques do not have to simplify in order to manipulate, do inference, etc. This is contrast to the human mind, which has a distinct limitation on the amount of detail it can keep track of simultaneously. There are two kinds of computational techniques that seem to be relevant in this case: local data mining and "descriptive" individual-based simulation.

Data mining and knowledge discovery aim to discover patterns in data using computer algorithms. This is not a "magic" technique producing knowledge out of nothing, but relies upon the application of domain knowledge for its effectiveness. That is, by applying assumptions that can be safely made about the nature of the data presented, the kinds of pattern that might be relevant, etc. and its goals (in terms of the kind of fit to the data wanted), the machine searches for the patterns of the kind specified that fit the data in the way specified.[6] Some of these approaches are essentially local, in the sense that they are not searching for a pattern that is common to all the data but is valid for a coherent subset of the data. For example, if the algorithm were looking for patterns of spending behaviour it might find one pattern valid in high-risk situations and another in very-low-risk situations but none in between. Thus the generalisation from data to pattern can be done in a manner that takes some aspects of the context into account. Such techniques could be used to try and suggest how the data is divided by context, rather than looking for weaker global patterns attributing the deviations to noise. The local patterns might then suggest complicated sets of hypothesis, each valid in its own specific scope.[7]

Unsurprisingly, given the topic, quite a few of the chapters use, discuss or take ideas from agent-based modelling. Goldspink (Chap. 4) makes suggestions for what an adequate agent-based model following an emergentist approach would follow; Elsenbroich (Chap. 5) critiques agent-based models for their individualist approach; Burgemeestre et al. (Chap. 6) take two approaches to norms from the world of multi-agent systems (of which agent-based modelling is a subset) and both Andrighetto et al. (Chap. 8) and Janssen and Ostrom (Chap. 9) exhibit specific agent-based models and their results.

However, perhaps the most promising avenues combine approaches. Horne (Chap. 6) argues for a combination of laboratory studies with agent-based modelling, as does Janssen and Ostrom (Chap. 9). Further than these, agent-based modelling seems ideal for incorporating elements that come from qualitative data, especially in terms of specifying or validating the behaviour of individual agents in the simulation.

[6] This is not to minimise the cleverness of the algorithms that have been developed in this field; the task of finding feasible and efficient ways of doing this for different kinds of pattern, data and goal is hard.

[7] I have not gone into the difficulties of this approach, which include compensating for the fact that one can easily over-fit models (see patterns that are not justified by the data and are due to noise) when cherry-picking among a data set.

Techniques for the storage, management and tracking of data could aid the collection of heterogeneous databases of qualitative and quantitative data, automatically checking for consistency, completeness and even suggesting links. Data mining of the kind described above could be used to try and detect the regions of the data that might coincide with what are considered kinds of context and hence help classify observations and thus help identify the triggers identified by Bicchieri (2005). The new suggestions, ideas and tools from complexity science do not change the social science landscape out of all recognition, but they do open up the possibility of new affordances and combinations of approaches and so enable the triple difficulties of multiple levels, highly dynamic systems and context dependency to be better handled.

Conclusions

The unavoidable inter-level, dynamic and context-dependent nature of normative behaviour poses severe problems for its study. The various chapters herein do not define them away or attribute them to something like "noise," but rather, in their different ways, identify and grapple with them. In these chapters, and elsewhere, one can detect new developments that hold out the hope of significant progress in understanding social norms and, hence, making a significant contribution to the understanding of our society around us. These will promiscuously combine different approaches: the new ones leaking from computer and complexity science but also well-established ones from the social sciences and other fields such as cognitive science in a variety of multi-pronged syntheses.

Acknowledgements Bruce Edmonds was supported by the EU 6th Framework grant 33841, EPSRC grant EP/H02171X/1 and the Manchester Metropolitan University Business School.

References

Bicchieri, C. (2005). *The grammar of society*. Cambridge: Cambridge University Press.

Birkin, M., Allan, R., Beckhofer, S., Buchan, I., Finch, J., Goble, C., et al. (2010). The elements of a computational infrastructure for social simulation. *Royal Society of London Philosophical Transactions A: Mathematical, Physical, and Engineering Sciences, 368*(1925), 3797–3812.

Conte, R., Edmonds, B., Moss, S., & Sawyer, R. K. (2001). Sociology and social theory in agent based social simulation: A symposium. *Computational and Mathematical Organization Theory, 7*(3), 183–205.

Sawyer, R. K. (2005). *Social emergence: Societies as complex systems*. New York, NY: Cambridge University Press.

Index

M. Xenitidou and B. Edmonds (eds.), *The Complexity of Social Norms*, Computational Social Sciences, DOI 10.1007/978-3-319-05308-0,