An interdisciplinary approach to the emergence and enforcement of norms of coordination and cooperation

Author(s):
Groeber, Patrick Christian

Publication Date:
2009

Permanent Link:
https://doi.org/10.3929/ethz-a-005899516

Rights / License:
In Copyright - Non-Commercial Use Permitted
An Interdisciplinary Approach to the Emergence and Enforcement of Norms of Coordination and Cooperation

A dissertation submitted to the
ETH ZURICH

for the degree of
DOCTOR OF SCIENCES

presented by
PATRICK CHRISTIAN GROEBER
Diplom-Mathematiker, Universität Bremen
born 23 June 1978
citizen of Germany

accepted on the recommendation of

Prof. Dr. Dr. Frank Schweitzer, examiner

Prof. Dr. Dirk Helbing, co-examiner

2009
# Contents

1 Introduction  
1.1 What is a Norm? .......................... 3  
1.2 Coordination Norms and Cooperation Norms ....................... 3  
1.3 Research Questions .......................... 6  
   1.3.1 Coordination Norms and Opinion Dynamics ............... 7  
   1.3.2 The Dark Figure of the Violation of Cooperation Norms .... 10

2 A Microfoundation of Social Influence  
2.1 The Underlying Concepts of Opinion Dynamics ....................... 15  
2.2 A Microfoundation of Social Influence based on Cognitive Dissonance ... 17  
   2.2.1 The Model .................................. 18  
   2.2.2 Identifying Coordination as the Underlying Process ............ 20  
   2.2.3 Myopic Best Response Dynamics .......................... 23  
2.3 Relation to Models of Opinion Dynamics .......................... 27  
   2.3.1 Bounded Confidence Models .......................... 28  
   2.3.2 Voter Models .................................. 30  
2.4 Combining Consonant and Dissonant Perceptions ....................... 35  
2.5 Concluding Remarks .......................... 41

3 An Application: Local Cultures in Clusters of Firms  
3.1 The Need for Local Cultures .......................... 44  
3.2 Local Cultures as Norms or Conventions .......................... 46
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Modelling the Emergence of Consensual behavior</td>
<td>49</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Consensus Formation</td>
<td>49</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Network Formation</td>
<td>51</td>
</tr>
<tr>
<td>3.4</td>
<td>Findings</td>
<td>53</td>
</tr>
<tr>
<td>3.5</td>
<td>Concluding Remarks</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>Cooperation Norms and the Dark Figure of Norm Violation</td>
<td>63</td>
</tr>
<tr>
<td>4.1</td>
<td>A Public Goods Game with Unobservable Actions</td>
<td>64</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Detection Probability as a Result of Interaction</td>
<td>65</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Modeling the Norm Targets and the Inspector</td>
<td>66</td>
</tr>
<tr>
<td>4.1.3</td>
<td>Rationality and Knowledge</td>
<td>69</td>
</tr>
<tr>
<td>4.2</td>
<td>Assumptions and Basic Properties</td>
<td>71</td>
</tr>
<tr>
<td>4.3</td>
<td>Equilibrium Properties for Observable Proportion of Norm Violators</td>
<td>75</td>
</tr>
<tr>
<td>4.4</td>
<td>Unobservable Proportion of Norm Violators</td>
<td>80</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Estimation of the Dark Field of Norm Violators</td>
<td>80</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Equilibrium Properties</td>
<td>82</td>
</tr>
<tr>
<td>4.4.3</td>
<td>An Inverse Self-Fulfilling Prophecy Effect</td>
<td>85</td>
</tr>
<tr>
<td>4.5</td>
<td>Robustness of the Results</td>
<td>91</td>
</tr>
<tr>
<td>4.6</td>
<td>Concluding Remarks</td>
<td>92</td>
</tr>
<tr>
<td>5</td>
<td>The Effect of Nescience: An Experimental Design</td>
<td>95</td>
</tr>
<tr>
<td>5.1</td>
<td>Why Experiments?</td>
<td>95</td>
</tr>
<tr>
<td>5.2</td>
<td>A Simplified Public Goods Game With Unobservable Actions</td>
<td>97</td>
</tr>
<tr>
<td>5.3</td>
<td>Analysis of the Experimental Design</td>
<td>101</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Properties of the Theoretical Dynamics</td>
<td>101</td>
</tr>
<tr>
<td>5.3.2</td>
<td>The Choice of Parameters</td>
<td>103</td>
</tr>
<tr>
<td>5.4</td>
<td>Statistical Analysis</td>
<td>108</td>
</tr>
<tr>
<td>5.4.1</td>
<td>The Inverse Self-Fulfilling Prophecy Effect</td>
<td>108</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Testing the Model Assumptions</td>
<td>110</td>
</tr>
<tr>
<td>5.5</td>
<td>Concluding Remarks</td>
<td>111</td>
</tr>
</tbody>
</table>
Abstract  Social norms are the cement of society (Elster, 1989). Although the definition of a norm differs between disciplines, they are undoubtedly relevant for all social sciences. Key topics of current research are the emergence of norms (Bendor and Swistak, 2001) and their enforcement (Fehr and Fischbacher, 2004; Fehr and Gächter, 2002). In this thesis, we address both topics with an interdisciplinary approach: Following the tradition of mathematical sociology (Coleman, 1964), we combine sociological and economic theory with mathematical modeling and analysis as well as computer simulations. Our starting point is the game theoretic classification of norms by Ullmann-Margalit (1977) who distinguishes between norms of coordination and cooperation.

The first type of norm is relevant when coordination towards one of many alternatives is the most beneficial outcome for all individuals. While there is no incentive to deviate from an established norm, it is not clear how it becomes established, in particular if the alternatives are equivalent ex ante. In the first part of the thesis, we link different models of opinion dynamics with coordination norms. We show that in these models, social influence between individuals can formally be identified as a coordination process within a general model based on the social psychological concept of cognitive dissonance (Festinger, 1957). We thereby provide a common basis for many models of opinion dynamics which used to be separated before. Further, we investigate the influence of interaction networks on the emergence of norms in the context of firm clusters. We extend the opinion dynamics model of Deffuant et al. (2000) by a feedback mechanism between the firms’ business practice and their interaction. Our analysis shows that this mechanism can foster consensual behavior (“local cultures”) compared to the original model and therefore make the emergence of a local culture more likely.

In the second part, the focus shifts to cooperation norms and their enforcement. They are instrumental to solving cooperation problems which are characterized by a state that is optimal from the community’s perspective but exhibits an incentive to deviate for a selfish individual. Hence, the challenge addressed by cooperation norms is to enforce the desired behavior. Different to previous research, we assume that norm violations are not directly observable and can only be detected by an external inspector. We investigate the influence of subjective beliefs about how many people adhere to a cooperation norm on the actual extent of norm violation in a population. Popitz (1968) claims that the nescience of the extent of norm violation enhances norm adherence: If people knew that many of their fellow citizens violate the norm, they would be less reluctant to follow suit. However, we argue that there is an opposite effect caused by the interaction between norm targets and inspector: A higher confidence in a moral world causes less normative behavior due to decreased inspection incentives. In order to validate this inverse self-fulfilling prophecy effect, we develop an experimental design based on a simplified version of our model.


Summary  Social norms influence human interactions in various ways and are therefore a central topic in the social sciences. Here, they are often instrumental to solving social dilemmas. However, the function of a norm does not explain its existence (Coleman, 1990). In fact, explaining the emergence of norms is one of the most challenging problems in the social sciences (Bendor and Swistak, 2001). Further, sanctions and punishment of norm violation constitute methods of norm enforcement. This can be performed by an actor concerned, a third party or a Hobbesian Leviathan. If punishment is costly, the enforcement of norms can be described as a second order collective good problem (Coleman, 1990; Heckathorn, 1989). Here, altruistic punishment is a possible solution (Fehr and Gächter, 2002) but requires that norm violation be observable. The anti doping norm or the norm to pay income tax are examples where detection of norm violators is a pivotal part of norm enforcement.

In this thesis, we contribute to the research on the emergence and enforcement of social norms of coordination and cooperation. We combine sociological and economic theory with mathematical modeling and analysis, computer simulations and an experimental design of the empirical validation of our results. In the following paragraphs, the different chapters of the thesis are summarized.

In Chapter 1 we motivate our research on the emergence and enforcement of social norms of coordination and cooperation. After discussing different norm concepts and defining a norm in the context of this thesis, we explain the distinction between coordination norms and cooperation norms based on the game theoretic classification of Ullmann-Margalit (1977). It is pointed out that different research questions arise for the two types of norms: The question of emergence is pivotal with respect to coordination norms while enforcement is of small importance as there is no incentive for an individual to deviate from such a norm once it has been established. This aspect is closely related to the problem of equilibrium selection in coordination games (Harsanyi and Selten, 1988; Samuelson, 1997; Young, 1998). In contrast, cooperation norms are not self-enforcing: Here, the behavior prescribed by the norm is optimal at the level of the whole society whereas it is beneficial for a selfish individual to violate the norm. Therefore, in the context of coordination norms, our research focuses on their enforcement. In particular, we emphasize that detection of norm violators is necessary for norm enforcement by a sanction mechanism. However, the anti-doping norm or the norm to pay taxes are examples where the main challenge is to detect norm violators. We then tie in with Popitz (1968) by asking whether the nescience of the extent of norm violators has an effect on norm compliance.

In the first part of the thesis, we focus on coordination norms. Here, Chapter 2 links this type of norms to consensus formation in models of opinion dynamics (e.g. Axelrod, 1997;
Deffuant et al., 2000; Hegselmann and Krause, 2002). In these models, Flache and Macy (2008) and Flache and Mäs (2008) identify the concept of social influence as one basic underlying mechanism. According to social influence, interaction between individuals leads to the adaptation of their respective opinions (Abelson, 1964; Brass et al., 1998; Kerr and Tindale, 2004; Strang and Soule, 1998). In this chapter, we provide a general framework of social influence in opinion dynamics models based on the socio psychological concept of cognitive dissonance (Festinger, 1957). We thereby assume that individuals strive to minimize cognitive dissonance resulting from different opinions compared to neighbor individuals in a given social network where the dissonance increases with opinion differences. This scenario can be formally represented as a game whose Nash equilibria correspond to consensus opinions for homogeneous interactions. As there is no cognitive dissonance in a consensus state, the general model exhibits the basic property of a coordination game. We can further show that the opinion dynamics models mentioned above can be represented as best response dynamics of this general for a particular choice of parameters. We also derive some basic properties of the best response dynamics for convex opinion sets that correspond to the assumption of social influence in the opinion dynamics models. Finally, we also allow that individuals may perceive consonance due to deviation from the opinions of other individuals they dislike. Depending on the parameters and in particular on the social network, the model can reproduce consensus, extreme opinions or a smooth opinion distribution in the population.

In Chapter 3 we investigate the emergence of a particular set of norms, namely local cultures in clusters of firms. In economic geography, local cultures with rules like “do not hire from competitors”, “exchange ideas freely” or “deliver only the highest quality” are a phenomenon characterizing clusters like Silicon Valley (IT), London (financial services) or Prato, Italy (textiles). Such local cultures are instrumental for firms to realize the benefits of their co-location through scale and specialization effects as well as positive externalities. So far, research concentrated on how local cultures look (e.g. Porter, 1990; Pyke et al., 1990; Saxenian, 1994)) and how their enforcement can be ensured (Hollandér, 1990; Kandel and Lazear, 1992) while their emergence is far less understood. We contribute to this topic by adjusting the opinion dynamics model of Deffuant et al. (2000) to investigate how interactions between firms influence the development of their business pratices. In particular, we incorporate an evolving cooperation network among the firms where two firms only interact if their counterpart’s business practice is similar enough compared to their own and to that of firms they previously interacted with. We show by means of computer simulations that compared to the model of Deffuant et al. (2000), this consideration of past interactions often fosters the emergence of a consensual business practice which provides a basis for a local culture. This holds in particular if the firms’ threshold
with respect to the tolerance of different business practice is low. Hence, we argue that the consideration of previous interaction partners and their business practice is beneficial for the firms in a cluster as it fosters the emergence of the rules in a local culture.

In the second part of the thesis, the focus is shifted to the enforcement of cooperation norms. Here, we investigate the influence of subjective beliefs about the extent of norm violation on normative behavior. Although altruistic punishment can solve the second order collective good problem of norm enforcement Fehr and Gächter (2002), it requires that norm violation be observable. The anti doping norm in sports or the norm to pay income tax are examples where the detection of norm violators is costly and conducted by an external inspection institution (the inspector) and where the actual extent of norm violation is unknown. Considerations about the unknown field of undetected deviance are not new to sociology. An example is the debate between strain theory (Merton, 1957) and labeling theory (Becker, 1963; Kitsuse, 1962; Lemert, 1967; Tittle, 1980, 1969). Further, Popitz (1968) argues that the nescience of the actual extent of norm violation may strengthen a norm: If it were revealed that a large proportion of the population violates the norm, norm targets would be less reluctant with respect to norm violation. However, a precise theoretical and formal description of the effect of the dark figure of norm violation on normative behavior is lacking. In Chapter 4, we present a mathematical model based on a public goods game to analyze the interaction between subjective beliefs about the dark field on one hand and delinquency and control on the other. In this model, the norm targets can either contribute to the public good (adhere to the norm) or not (violate the norm). Further, an external inspector performs controls and receives a reward that increases with the number of detected norm violators. The probability of detection in turn depends on a norm violator’s investment in the concealment of her behavior and the frequency and quality of controls by the inspector. Further, types of actors are boundedly rational and aware of each other’s payoff functions. Although the actual extent of norm violation is not observable, there is usually information about the number of detected norm violations. This information is then combined by the norm targets and the inspector with their external subjective beliefs about norm adherence to an estimation of the actual extent of norm violation.

We then analyze the asymptotic behavior of the model and in particular the extent of norm violation in the population in equilibrium. As a result, we show that multiple equilibria are possible when there is a dark figure of norm violation whereas there is a unique equilibrium if the proportion of norm violators is observable. Hence, the extent of norm violation in equilibrium may depend on the initial state of the system in the first case. Further, we identify an inverse self-fulfilling prophecy effect (cf. Merton, 1957): The higher the suspiciousness by the norm targets and the inspector towards norm violation, the lower
the actual extent of norm violation in equilibrium usually is. Hence, a change in beliefs has the opposite behavioral effect. The effect of the interaction between norm targets and inspector is therefore diametrically opposed to the preventive effect of nescience described by Popitz (1968).

In order to validate our model and in particular the inverse self-fulfilling prophecy effect, we suggest an experimental design in Chapter 5. After arguing why laboratory experiments are preferable to other methodologies with respect to a validation, we present a design based on our model of the previous chapter. However, we use a simplification of our general model in order to reduce the complexity of the participant’s decision. Here, we provide a detailed description of experiment with particular emphasis on what information is available to the participants. Although our theoretical predictions do not depend on the model parameters, we discuss which parameter setting enhances the observability of our results in the experiment based on a theoretical analysis and computer simulations of the simplified model. Further, we specify adequate methodologies and measurements to verify our hypotheses and model assumptions in the experiment.

Finally, Chapter 6 concludes this thesis with a discussion of the results and possible extensions for future research.
Acknowledgements  I would like to thank my advisor Frank Schweitzer for his invaluable support and help, and for giving me freedom in my research at the same time. Moreover, this thesis rests on the team work with several other people. I would like to thank Kerstin Press and Heiko Rauhut for the great collaboration and sustained support and reliability. I enormously profited from our numerous discussions and their helpful advice with respect to economic and sociological theory.

Working at the Chair of Systems Design at ETH Zurich has been a pleasure, and I would like all the people who made these years enjoyable. Special thanks are due to Matthias Feiler and Jan Lorenz who provided me with detailed comments and suggestions, in particular with respect to mathematical problems. I profited very much from the work of Markus Geipel on graph visualization. I owe profound thanks to Michael König who always had an ear for me and pointed me to new literature and interesting courses more than once. I further thank Jaba Ghonghadze and Claudio Tessone for useful comments on parts of the manuscript.

Moreover, I am grateful to Ulrich Krause for his excellent mentoring at the University of Bremen, his encouragement to start this dissertation project and inspiring discussions about science and beyond. I want to thank María Sáez-Martí for her excellent courses on game theory and the possibility to present my work at the Ph.D. seminar of the Institute of Empirical Research in Economics at the University of Zurich. I further benefited from comments by Andreas Diekmann, Andreas Flache, Rainer Hegselmann, Dirk Helbing, Wander Jager, Michael Mäs, Volker Müller-Benedict and Thomas Voss.

I would also like to thank the participants of the “Frühjahrstagung 2007 der Sektion Modellbildung und Simulation der Deutschen Gesellschaft für Soziologie” in Kiel and the International Conference on Rational Choice and Social Institutions in Zurich, 2007. Their insightful comments were useful in further improving this dissertation.

It would have been very difficult to accomplish this work without the support of my friends and, in particular, my parents who were a great help during my relocation to Zurich. Finally, I want to thank Kerstin Press for being not only a splendid scientific advisor, but also an understanding partner and friend.

Patrick Christian Groeber
Zurich, June 2009
Chapter 1

Introduction

Recently, there is a growing interest in interdisciplinary research. Its objectives include to “answer complex questions”, “address broad issues” and “solving problems that are beyond the scope of any one discipline” (Klein, 1990, p. 11). Social norms are intrinsically adequate for an interdisciplinary approach as they govern human behavior in various complex ways. Actions may be prescribed or proscribed by formal norms as laws, but also informal norms, i.e. unwritten consensual agreements in a society or a subgroup of individuals. Therefore, social norms are the cement of society (Elster, 1989). Their enforcement is a crucial asset for the stabilization of social order which is one of the core topics in sociology (Parsons, 1937). However, the importance of social norms is not restricted to sociology: “No concept is invoked more often by social scientists in the explanation of human behavior than ‘norm’” (Gibbs, 1968, p. 212). Although traditional economics mostly disregarded the concept and even developed a “norm against norms” (Eggertsson, 2001, p. 76), it gained in importance after the rise of new institutional economics (Coase, 1937, 1960; Eggertsson, 1990; Nelson and Winter, 1982; North, 1986; Williamson, 1985). Today, norms are invoked to explain a variety of economic phenomena including innovation, economic development as well as inter-firm relations.

Further, one can recognize an increased use of mathematics in the social sciences. Although it is more prominent in econmics, mathematics is also used in sociology where Lazarsfeld (1954) and most notably Coleman (1964) are early contributions. Its strongest impact on sociology is by providing statistical tools, e.g. for network analysis (Wasserman and Faust, 1994), event-history analysis (Blossfeld and Rohwer, 2002) or hierarchical linear modeling (Snijders and Bosker, 1999). However, as Edling (2002, p. 202) points out, “the use of mathematics in sociology is not about giving a quantitative approach to data preference over a qualitative approach”. Mathematics are also used in the form of theoretical models of social phenomena to generate hypotheses and check the consistency of sociological
Chapter 1. Introduction

theories (Carley, 1997). The mathematical analysis of these models can be complemented by computer simulations. In the following, we point out how mathematical modeling can contribute to our understanding of the emergence and enforcement of social norms.

The fact that norms are instrumental to solving social dilemmas does not explain their existence:

“Much sociological theory takes norms as given and proceeds to examine individual behavior or the behalf of social systems when norms exist. Yet to do this without raising at some point the question of why and how norms come into existence is to forsake the more important sociological problem in order to address the less important.” (Coleman, 1990, p. 241)

It is still an open problem how social norms emerge. Moreover, “the question, Why are there norms? is perhaps one of the most fundamental problems that the social sciences have ever tackled.” (Bendor and Swistak, 2001, p. 1493).

Further, the enforcement of norms is problematic. It usually proceeds through the deterring effect of punishment of norm violators. This can be described as a “second order collective good problem” (Coleman, 1990; Heckathorn, 1989): If punishment of norm violation is costly and does not imply a monetary reward for the executor, there is no incentive for sanctioning deviant behavior. Still, norms requiring enforcement exist in all kinds of human societies (Henrich et al., 2001), suggesting that punishment of norm violators does occur. One possible explanation of this is altruistic punishment (Fehr and Gächter, 2002). In general, punishment can be performed by an actor affected by the norm violation, by a third party (Fehr and Fischbacher, 2004) or a Hobbesian Leviathan. For any punishment to occur, the norm violation has to be detected. This detection is often taken for granted, but may itself be costly and thereby constitute a second order public good. Furthermore, there are examples where detection is a competition between norm violators and inspection institutions. Such settings include the norm to pay taxes or the anti-doping norm. As a consequence, not all norm violations may be detected resulting in a dark figure of norm violation. We argue that this dark figure and its effect on the “moral confidence” of the actors, i.e. their subjective beliefs of the actual extent of norm violation, may influence their decision whether to adhere to the norm or not. Put differently, “detection influences defection”.

Before we elaborate in detail the research questions addressed in this thesis, it is first clarified how we define a “norm”.

---

1.1 What is a Norm?

The social science literature provides various definitions, terminologies and theories with respect to norms. An overview and discussion of many norm definitions is provided by Gibbs (1965), Biddle and Thomas (1966), and Rommetveit (1953). From the perspective of sociology, Opp (2001a,b) distinguishes two types of definitions:\(^2\) *oughtness definitions* and *behavioral definitions*. The first type defines a norm as an expectation by individuals that a particular behavior, belief or attitude ought to be performed or held (e.g. Homans, 1974). Violation of an oughtness norm often entails internal sanctioning as guilt or shame. For behavioral norms, moral considerations are irrelevant. Here, norms are regularities of actual behavior where external sanctioning occurs with non-zero probability if the respective behavior is not performed (Axelrod, 1986; Calvert, 1995; Rutherford, 1996).

In this thesis, we do not commit to one of the above types of norm definitions. Instead, we assume that a norm can be constituted by both oughtness and behavioral aspects whereas the focus depends on the considered norm and in particular on whether there are intrinsic incentives for norm violation. If there are such incentives, the degree to which actors feel obliged not to violate the norm is crucial as it determines the extent of normative behavior in a population. On the other hand, oughtness becomes less important if it is in an actor’s best interest not to violate the norm. As we point out in the following section, norms have been classified with respect to these individual incentives for adherence to the norm.

1.2 Coordination Norms and Cooperation Norms

Consider the following examples of human behavior (partly adopted from Voss, 2001, p. 106):

1. In continental Europe, car drivers use the right hand side of the road (Young, 1998, p. 16 ff).

2. Most persons use computer keyboards with the “QWERTY” layout (David, 1985).

3. In many high school classes, those students who learn eagerly are not popular. Similarly, in working groups in large organizations, workers regularly restrict their output (e.g. Homans, 1950), and overperformers are targets of informal sanctions.

Chapter 1. Introduction

4. In the community of Internet users, a rule prohibiting the use of e-mail for commercial advertising is widely respected (Fukuyama, 1995, p. 196).

5. In sports contests, athletes are expected to respect the rule prohibiting the use of drugs that enhance athletic performance and thus might improve their chances of winning a prize (Bird and Wagner, 1997).

We can identify social norms in all the above examples. However, there are substantial differences between the norms: In the first example, a car driver faces the decision whether to drive on the left or on the right hand side. In continental Europe of the Twenty-First Century, the driver should obviously choose the right hand side in order not to be fined or to endanger himself or other people. Hence, there is no incentive to violate the norm. However, this holds due to the common expectation about the behavior of other car drivers. In general, the interaction between the car drivers has the structure of a coordination game. For two players, it can be illustrated as follows:

\[
\begin{array}{c|c|c}
 & \text{left} & \text{right} \\
\hline
\text{left} & 1,1 & 0,0 \\
\text{right} & 0,0 & 1,1 \\
\end{array}
\]

The preferred outcome for all actors is to choose identical strategies (i.e. drive on the same side) whereas they are indifferent towards which side is actually chosen. Both preferred outcomes are Nash equilibria, i.e. in these states neither player can improve by unilateral deviation. However, the players face a coordination problem in case there are no shared expectations about which equilibrium should be chosen.

The second example exhibits a similar structure. If everybody else uses the QWERTY-layout, there is no incentive to learn and use a different layout. However, as David (1985) points out, there is an alternative layout named DSK (Dvorak Simplified Keyboard) which is based on motion studies of typists’ hand movements and which is considered to be superior to QWERTY. Nevertheless, its superiority only pays off for an individual if it is used by a sufficient number of people. This structure is captured by the typewriter game (Young, 1998, p. 26):

\[
\begin{array}{c|c|c}
 & \text{QWERTY} & \text{DSK} \\
\hline
\text{QWERTY} & 4,4 & 0,0 \\
\text{DSK} & 0,0 & 5,5 \\
\end{array}
\]

See also Helbing (1991) with respect to the coordination of pedestrians.

The left number in the game matrix denotes the row player’s payoff, the right number the column player’s payoff.
Again, choosing identical layouts is a Nash equilibrium. With respect to the equilibria, choosing DSK is more beneficial for all players. However, QWERTY evolved to today’s standard layout. According to David (1985, p. 332), this “historical accident” can be explained by the stochastic and path-dependent selection process that favored QWERTY due to an initial advantage. Finally, the process locked in the actually inferior state where there is no incentive for users to change the layout unilaterally.

The remaining examples (3-5) reflect a different type of norm. Here, there is an incentive to violate the norm: Students and workers would profit from increased effort by better grades or higher salaries respectively, commercial advertising in emails is beneficial for the sender, and drug use enhances an athlete’s performance and therefore increases the probability of a higher prize money. However, norm violation would be at the expense of the remaining agents: Students and workers would also have to increase their effort to maintain their grades or income or take a loss. Internet users would be disturbed by commercial emails, and other athletes’ winning probability is reduced. Therefore, the norm is beneficial for all agents in the sense that it shall prevent selfish actions at the expense of the community by the threat of formal (example 5) or informal (examples 3,4) sanctions. If only two actors are involved, the structure of these situations can be exemplified by the prisoner’s dilemma (e.g. Poundstone, 1992; Rapoport and Chammah, 1965):

\[
\begin{array}{cc}
C & D \\
C & 3,3 & 0,5 \\
D & 5,0 & 1,1 \\
\end{array}
\]

Here, “C” (cooperate) denotes norm adherence while “D” (defect) represents norm violation. The only Nash equilibrium is defection by the players whereas both could improve if they cooperated. Therefore, cooperation constitutes the desired state.

The discussion of the examples corresponds to the game theoretic classification of norms by Ullmann-Margalit (1977) which is also discussed by Voss (2001). The first two examples constitute coordination norms (or conventions according to Lewis (1969)) that are instrumental to solving coordination problems. Coordination norms are self-stabilizing and self-enforcing: Once it is established, there is no incentive for norm violation both on the individual and on the community level. Further, coordination norms are often unplanned: Uncoordinated individual actions performed to reach individual goals lead to their emergence. Hence, a behavioral regularity arises first and finally becomes a norm (Horne, 2001a,b; Opp, 1982; Sugden, 1998; Sumner, 1906). The remaining examples (3-5) are instances of prisoner’s dilemma norms (Ullmann-Margalit, 1977) which we also refer

---

5 Additionally, there is a Nash equilibrium in mixed strategies.
to as *cooperation norms*. The function of these norms is to establish a behavior that is beneficial to the community but not self-enforcing as there are incentives for unilateral deviance. Thus, there is a need for sanctioning norm violation. As the sanctioning of norm violators is usually costly, cooperation norms are second-order public goods which are instrumental to providing first-order public goods (Coleman, 1990; Heckathorn, 1989; Oliver, 1980).

In the following, we identify the research questions addressed to in this thesis with respect to the emergence and enforcement of coordination and cooperation norms.

### 1.3 Research Questions

As we pointed out in the previous section, there is a “demand” for norms of coordination and cooperation (Coleman, 1990), i.e. the norms are instrumental to solving coordination problems or dilemma situations. However, a demand for a norm does not imply that the norm will emerge. Hence, this functional approach does not explain “what mechanisms contribute to the enforcement, or ‘effective realization’, of norms” (Voss, 2001, p. 105). With respect to a coordination problem, it is of particular interest whether one observes a behavioral regularity (or even identical behavior among the actors) that finally becomes a convention and, if so, which behavior. As conventions are self-enforcing, norm enforcement is usually not an issue here. This is different for cooperation problems: As there is an incentive for the actors to deviate from the behavior that is optimal on the community level, norm enforcement and the extent of compliance to the desired behavior is a central issue.

Due to the intrinsic differences between coordination norms and cooperation norms, the thesis can be divided in two parts. In the first part, Chapters 2 and 3 cover coordination norms and investigate in particular the relation between the emergence of this type of norms and a class of processes usually referred to as *opinion dynamics*. The second part, Chapters 4 and 5, covers cooperation norms. Here, unlike the existing literature, we account for the fact that the actual extent of norm violation may be unknown, i.e. that there is a *dark figure* of norm violation. We then investigate how subjective beliefs about this dark figure feed back to the actual extent of norm violation. In the following, we expose the respective research questions addressed to in the two parts more.

---

Ullmann-Margalit (1977) also identifies *norms of partiality*. This class of norms requires an asymmetry with respect to the actors in the sense that some are more privileged than others.
Figure 1.1: Thesis Structure: Chapters 2 and 3 investigate the emergence of coordination norms while Chapters 4 and 5 examine the enforcement of cooperation norms.

1.3.1 Coordination Norms and Opinion Dynamics

Coordination problems can be represented in the context of game theory (Ullmann-Margalit, 1977; Voss, 2001). In a coordination game, the basic challenge for individuals is to coordinate on one of the possibly many equilibria. Assuming perfectly rational individuals, one possible solution is the concept of Harsanyi and Selten (1988) which selects an equilibrium for any finite game (i.e. a game with a finite number of players and a finite strategy space). Apart from the fact that this concept is not a unique solution to the problem, its main disadvantage is that the assumption of perfect rationality is unrealistic in human decision making (Aumann, 1997; Gigerenzer and Selten, 2001; Kahneman, 2003; Kahneman and Tversky, 1979; Simon, 1957). Further, as an equilibrium concept, it does not consider which processes actually lead to an equilibrium.

In contrast, evolutionary game theory (Hofbauer and Sigmund, 1988; Maynard Smith, 1982; Weibull, 1995) abandons the assumption of strategic interaction between the individuals. Here, actors are “programmed” to play a certain strategy and reproduce according to their “fitness” which depends on the strategy’s average payoff for playing the base game according to random pairwise matching in the population. This induces an evolution of the frequency of a strategy that corresponds to the share of the population playing that strategy. The most prominent example of such dynamical process induced by the underlying game are replicator dynamics (Taylor and Jonker, 1978). For a discussion of similar
evolutionary processes, see Weibull (1995, chapter 4).\textsuperscript{7} In general, the long-term behavior of the system (i.e. with respect to whether a particular strategy becomes extinct or not) depends on the initial frequencies of the strategies. Further, evolutionary game theory provides a refinement of the concept of Nash equilibrium, namely \textit{evolutionary stability} (Maynard Smith, 1974; Maynard Smith and Price, 1973). Simply said, an evolutionarily stable strategy is one that cannot be invaded by a sufficiently small share of other “mutant” strategies.

However, although evolutionary game theory resigns from the assumption of hyper-rational actors, it basically switches to the opposite extreme by assuming that actors do not make rational considerations at all. The truth lies probably in between. With respect to the evolution of behaviors and behavioral expectations that finally lead to the emergence of conventions, it is often assumed that the actor’s decisions are influenced by their bounded knowledge of past experiences (Axelrod, 1986; Bicchieri, 1990; Lewis, 1969; Sugden, 1986; Wärneryd, 1990). In the models of Young (1993) and Kandori \textit{et al.} (1993), which investigate the evolution of conventions, the actors usually choose a best response with respect to their experience based on a limited history of previous plays, also referred to as \textit{fictitious play} (Brown, 1951; Fudenberg and Levine, 1998). However, with a small probability, they make a random choice which represents that the actors occasionally experiment with different strategies or simply make mistakes. This defines a Markov chain of the set of possible histories. Further, it is investigated which states occur with a positive probability if the error probability tends to zero. Such states are called \textit{stochastically stable} (Foster and Young, 1990). Stochastic stability constitutes a further refinement of the equilibrium concept as this definition is independent of the initial state of the process.\textsuperscript{8} While for $2 \times 2$ games, these equilibria coincide with risk dominant equilibria (Harsanyi and Selten, 1988), these concepts differ for general games (Young, 1993). Further, the effect of local interactions in similar models is investigated by Ellison (1993) and Blume (1995).\textsuperscript{9}

While the above models define dynamical processes of human behavior based on games and in particular coordination games, another type of process modeling social interaction seems unrelated to the emergence of coordination norms at first sight as they are not explicitly based on games: Models of opinion dynamics reflect the evolution of opinions, behaviors, beliefs or attitudes (in the following only referred to as “opinions”). These stylized models are used to study collective phenomena as finding or not finding consensus (Deffuant \textit{et al.}, 2000; Hegselmann and Krause, 2002), minority opinion spreading (Galam, \textsuperscript{7}For a microfoundation of several types of evolutionary dynamics, see Helbing (1996).
\textsuperscript{8}For a positive error probability, the induced Markov chain is ergodic.
\textsuperscript{9}See also Helbing (1994) for a model that combines random behavioral changes with interaction effects which lead to imitative, avoidance or compromising processes.
1.3. Research Questions

2002; Tessone et al., 2004) and the emergence of political parties (Ben-Naim, 2005) or extremism (Deffuant et al., 2002). Real world applications of opinion dynamics include “wisdom of crowds” (Surowiecki, 2004) due to “swarm intelligence” (Bonabeau et al., 1999) or imperfect prediction markets (Chen et al., 2004), decision making in committees of experts (Visser and Swank, 2007) or the spread of participation processes in democratic societies (de Sousa Santos, 1998).\textsuperscript{10}

In models of opinion dynamics, an opinion usually has no intrinsic value but evolves according to prescribed rules of social interaction. These rules differ between models as well as the set of possible opinions or the underlying interaction networks. However, we can often identify a main property of a coordination process: In many models (e.g. Axelrod, 1997; Deffuant et al., 2000; Hegselmann and Krause, 2002), consensus among the actors (i.e. all actors exhibit identical opinions) is a fixed point of the respective dynamics. With respect to a coordination process, this reflects that there is no incentive to deviate from a state of unanimity once it has been established. Additionally, there may be other fixed points where different subgroups of the population have different opinions.

Although there is an indication that these models may be related to evolutionary coordination processes, a formal derivation of this conjecture is lacking. In Chapter 2, we fill this gap by providing a microfoundation of social influence in models of opinion dynamics. We introduce a general model based on the social psychological concept of cognitive dissonance (Festinger, 1957) and show that it can be interpreted as a coordination game. Further, we show that common models of opinion dynamics (Axelrod, 1997; Deffuant et al., 2000; Hegselmann and Krause, 2002) can be represented as best response dynamics resulting from the general model. Hence, we formally identify these models as coordination processes and provide a mutual basis for models that differ in many aspects.

In Chapter 3, we use an opinion dynamics model to investigate the emergence of a particular set of norms, namely rules on acceptable business practice among firms in a cluster, so-called local cultures. Co-location may convey benefits to cluster firms (Marshall, 1920), and local cultures are instrumental to guaranteeing that these benefits are not compensated by friction loss due to different business practices. Further, local cultures cannot be definitely classified as coordination norms or cooperation norms: In general, it is not clear whether there is an incentive to deviate from common business practice. There are examples for “cooperative” local cultures as well as for “defective” ones, and neither type has proven itself to be superior. Therefore, we assume that coordination on any business practice is beneficial for firms. Models of opinion dynamics have been used to study the emergence of such a consensus. As we point out, in particular the model of Deffuant et al.

\textsuperscript{10}For a more detailed overview of applications of opinion dynamics, see Lorenz (2007b, Section 2.1).
(2000) reflects basic properties of the interaction between firms. Further, social networks influence the emergence of norms (Coleman, 1990; Ellickson, 1991; Opp, 2001a; Putnam, 1993; Taylor, 1996; Voss, 1998). We then propose an extension of the model with respect to how firms select their interaction partners and investigate which mechanism is more beneficial, i.e. which mechanism is more likely to lead to a local culture.

1.3.2 The Dark Figure of the Violation of Cooperation Norms

Although every society has its normative standards, we perfectly know that not everybody adheres to all social norms. Most of us already experienced situations, where we “just happened” to break the rules. Some individuals are detected, but there is certainly a dark field of undetected deviance. While it is tempting to be in the position to know every indecency, it is questionable that a society would indeed be better off if their members had such a complete knowledge. William Makepeace Thackeray described in his classical novel some potential horrors of completely unravelling misconduct in society:

“Just picture to yourself everybody who does wrong being found out, and punished accordingly. Fancy all the boys in all the school being whipped; and then the assistants, and then the headmaster ... Fancy the provost marshal being tied up, having previously superintended the correction of the whole army. ... The butchery is too horrible. The hand drops powerless, appalled at the quantity of birch which it must cut and brandish. I am glad we are not all found out.” Thackeray (1869)

Detection of norm violation is a necessary condition for the enforcement of social norms. When detection and punishment is costly, enforcement can be described as a “second order collective good problem” (Coleman, 1990; Heckathorn, 1989). Recently, there has been a growing interest among sociologists, economists and social psychologists to uncover the mechanisms of norm enforcement and human motives for cooperation and punishment. It was revealed that humans have strong emotions of envy when left as suckers in a dilemma and of guilt when they outsmart their opponents. These emotions let even strangers exert high altruistic efforts to punish the uncooperative – when punishment is available (Fehr and Gächter, 2000; Ostrom et al., 1992; Yamagishi, 1986). Thus, altruistic punishment became a core interdisciplinary subject since the article of Fehr and Gächter (2002), in which they coined the term altruistic punishment. No wonder, the discovery of such altruistic motives is considered as one of the major breakthroughs in understanding cooperative relations of modern life (as documented in the reviews of Camerer, 2003; Camerer and Fehr, 2006; Fehr and Gintis, 2007; Nowak, 2006; Sigmund, 2007).
1.3. Research Questions

However, while there is a growing interest in the punishment problem, there is little research on the detection problem, which captures second order collective goods likewise (see Rauhut and Krumpal, 2008, for a detailed discussion). There are many examples where detection is the crucial problem. In the case of income taxes, we are neither able to observe whether one particular individual truthfully declares her income, nor do we know the overall extent of tax evasion without inspection. Similar examples are doping in sports and fare dodging in public transport. Often the detection problem is accompanied with external inspections. Here, norm enforcement is not performed by the targets of the norm themselves but by external “inspectors”. However, in virtually all experiments on the enforcement of social norms, norm violation is always observable, the rule breakers are known. Such experiments focus on the conditions, in which humans are willing to punish norm violators at own costs, that is, altruistic punishment (reviewed in Camerer, 2003; Fehr and Gintis, 2007). One exception, focusing on the unknown rate of norm violations, is the laboratory experiment by Rauhut (2009).

The importance of detection and the so-called “dark figure” of norm violation due to imperfect detection has been emphasized in social research early onwards. In particular, the dark figure is important for understanding the relationship between social class and norm violation. Strain theory (Merton, 1957) assumes that most members of society have similar goals, however dissimilar means for reaching these goals. As a consequence, lower class members will seek illegitimate ways for obtaining these goals more often than higher class members. While empirical tests with direct correlations between social class and norm violations revealed indeed such a negative relationship, it is argued in “labeling theory” that the dark field covers larger parts of higher than of lower classes: In the case of crime, members of higher classes are less often suspect of crime, less often criminalized, less often arrested, accused, prosecuted and punished. Such a bias in the so-called “filtering process” of crime produces spurious correlations between social class and norm violations (Becker, 1963; Jennes, 2004; Kitsuse, 1962; Lemert, 1967; Mehlkop and Becker, 2004; Paternoster and Iovanni, 1989; Tittle, 1980, 1969).

\(^{11}\) According to Coleman (1990), we use “target of the norm” as an expression for the actors whose behavior is subject to the normative expectation. Note that we use “norm targets” as a short version for target of the norm.

\(^{12}\) In this experiment, norm violators are not known to norm enforcers and the impact of the beliefs of norm targets and inspectors about the actual, detected and non-detected norm violations on normative and control behavior is investigated.

\(^{13}\) We define the dark figure as the total number of undetected norm violations. Note that we use dark figure and dark field synonymously.

\(^{14}\) We use the term “norm violation” and “crime” synonymously and regard it as non-cooperative behavior in terms of contributing to public goods in the tradition of Coleman (1990).
Chapter 1. Introduction

The focus of labeling theory on detection captures only one part of the problem. To put it simple, it is restricted to the effect on observed norm violations: If the inspectors believe that a particular subgroup of the population is more likely to violate the norm, more norm violations will be revealed for this subgroup, while the actual extent of norm violations remains unaffected.\textsuperscript{15} On the other hand, we know from social constructionism, in particular from the Thomas theorem (Thomas and Thomas, 1928, p. 572) that “if men define situations as real, they are real in their consequences” (cf. also Merton, 1995). A specific version of the Thomas theorem is the well known “self-fulfilling prophecy” (Merton, 1957). Further, Popitz (1968) elaborated a theory how beliefs of the extent of unknown norm violations influence normative behavior. Herewith, he highlighted the ambiguity of detection in a meanwhile classical, though German, text. While revelation of norm violations is a precondition of punishment and deterrence, the knowledge that others violate the norm, destabilizes the normative system anew.

“The public affair [of a detected norm violation] reveals unambiguously that the norm is more sensitive and vulnerable than publicly perceived. . . . The hypocrisy of the public indignation . . . is conducive in so far as it supports the sacredness and splendor of the norm.” (p. 14). “The punishment is only effective as long as the majority does not ‘get what it deserves’. The preventive effect of punishment only remains if the general prevention of the dark figure persists. The splendor and misery of punishment rests upon ‘the marvelous, the beauty nurturance of nature’, which ensures that we know nothing — or at least very little.” (p. 20) Popitz (1968, translated by the authors)

Hence, if too many norm violators are detected, the prevalence of the norm is weakened. Moreover, punishment can be prone to such paradoxical effects as well. Thus both, too much and too little detection and punishment can weaken the social norm:

“If the neighbor to the right and to the left is punished, the punishment loses its moral power. Something that happens to almost everybody is no longer discriminating. Even punishment can wear out. If the norm is enforced too rarely, it looses its teeth, if it is enforced too often, the teeth deaden. . . . If the neighbor to the right and to the left is punished, it becomes conceivably evident that the neighbor does not comply with the norm. If too many are pilloried, the pillory is losing its horrors, the rule breaking loses its abnormality

\textsuperscript{15}Note that Lemert (1967) formulated the idea of a feedback loop between inspectors’ beliefs and norm violations by introducing the terms “primary” and “secondary deviation”.
1.3. Research Questions

and the perception that something has been “broken” vanishes.” Popitz (1968, p. 17, translated by the authors)

Whereas the investigation of the relationship between social class position and coverage of the dark field has received great attention, the precise theoretical mechanisms are still unknown. More specifically, we lack a precise theory how subjective beliefs about the dark field of norm violations on the one hand and normative and control behavior on the other interact with each other. So far, we have two theories, which focus only on one side of the medal. Labeling theory emphasizes the inspectors’ side, while Popitz (1968) emphasizes the side of the norm targets. The game theoretic approach developed by Tsebelis (1989, 1990) takes interaction between norm targets and inspectors into account. However, the interaction specified by Tsebelis (1990) neglects a dark figure because here, perfectly rational players can foresee their opponents’ actions and therewith the detected and undetected extent of norm violation.

In Chapter 4, we develop a mathematical model which incorporates those three approaches above. We analyze the impact of the dark figure on the interaction between norm targets and inspectors. We can show that the interaction between norm targets and inspectors reverts the self-fulfilling prophecy effect: In most cases, it is not true that a higher confidence in a moral world reinforces normative behavior. In contrast, a higher confidence in a moral world causes less normative behavior due to decreased inspection incentives.

An empirical validation of our model and in particular of the inverse self-fulfilling prophecy effect is discussed in Chapter 5. We point out that experiments are more suitable for the validation of our results compared to data analysis or survey studies. Further, we discuss an experimental design based on the extended public goods game in Section 4.1. While modifications were conducted to keep the actions for the participants of a future experiment as simple as possible, the basic properties remain unchanged. Finally, we analyze which parameter settings may foster the observability of our results in an experiment.
Chapter 2

A Microfoundation of Social Influence

In this chapter, we provide a microfoundation of the social influence between interacting agents based on the social psychological concept of cognitive dissonance (Festinger, 1957; Harmon-Jones and Mills, 1999). It assumes that agents choose the opinion that minimizes cognitive dissonance with respect to given opinions of other agents. We show that this mechanism can be interpreted as a process of coordination among the interacting individuals. Further, we highlight how common models of opinion dynamics can be interpreted as special instances of our model. Finally, we incorporate the concept of rejection into our approach. We show by means of computer simulations how a combination of social influence and rejection, different to existing models, can lead to a smooth and multimodal (quasi-)stationary distribution of opinions in the population.

2.1 The Underlying Concepts of Opinion Dynamics

Following Flache and Macy (2008) and Flache and Mäs (2008), one can identify two basic social psychological mechanisms in many models of opinion dynamics: social influence and homophily. According to social influence, interaction between individuals leads to the adaptation of their respective opinions (Abelson, 1964; Brass et al., 1998; Kerr and Tindale, 2004; Strang and Soule, 1998). Here, interaction takes place in a social network of individuals where agents interact locally with adjacent individuals (their neighbors). The implementation of social influence in an opinion dynamics model usually depends on the set of possible opinions. In the model of Axelrod (1997) an opinion is a discrete vector representing an agent’s attitudes to different aspects, and an agent randomly adopts a
component of her counterpart’s opinion during interaction. In the voter model introduced by Holley and Liggett (1975) opinions are binary, and agents choose an opinion with a probability that corresponds to its frequency in the agent’s neighborhood. If opinions are continuous (e.g. Abelson, 1964; Deffuant et al., 2000; Krause, 2000), i.e. real numbers or real vectors, adaptation is usually attained by a (weighted) average of the opinions of interacting agents. Further, the implementations of social influence differ with respect to the modality of interaction: The agents may either interact simultaneously or sequentially with one or several of their neighbors. For a fixed network, such social influence leads to consensus, i.e. all agents finally exhibit identical opinions, if the network is connected (Abelson, 1964; French, 1956; Harary, 1959). However, social networks typically emerge out of the interaction of individuals (Granovetter, 1985). According to the concept of homophily, more similar agents interact more frequently (Byrne, 1971; Kandel, 1978; Lazarsfeld and Merton, 1954; McPherson et al., 2001; Rogers and Bhowmik, 1970). Axelrod (1997) incorporated homophily in his model and showed by means of computer simulations that this may result in local convergence and global diversity of opinions: By allowing interaction between adjacent agents only if their opinions are similar enough, the interaction network constantly changes. On the one hand, homophily reinforces the effect of social influence as similarity of opinions leads to more interaction and interaction increases similarity. But on the other hand, it inhibits the interaction between agents with sufficiently different opinions. This finally leads to clusters of agents with identical opinions whereas diversity is not eliminated as the opinions in distinct clusters may differ. A similar mechanism of homophily in models of continuous opinion dynamics is bounded confidence (Hegselmann and Krause, 2002; Lorenz, 2007a). Here, the interaction network is determined by the agents’ opinions: Two agents are adjacent if their opinions are sufficiently similar.

Recent research in opinion dynamics has investigated the effect of different network structures (Fortunato, 2004, 2005; Stauffer and Meyer-Ortmanns, 2005; Suchecki et al., 2005; Weisbuch, 2004), agent heterogeneity with respect to their level of homophily (Deffuant et al., 2002; Lorenz, 2008; Weisbuch et al., 2005, 2002) and a co-evolution of opinions and network (Centola et al., 2007; Vazquez et al., 2008). However, following Flache and Mäs (2008), the concepts of social influence and homophily are not sufficient to explain why opinions sometimes drift away from moderate to extreme positions (cf. Earley (2000) in the context of team cultures). Therefore, some models (Fent et al., 2007; Flache and Macy, 2006; Jager and Amblard, 2005; Kitts, 2006; Salzarulo, 2006) incorporate the complementary concepts of rejection and heterophobia. According to rejection, individuals change their opinions to become more dissimilar to interaction partners they do not like.

---

1 Our local cultures model in Chapter 3 also exhibits a co-evolution of opinions and interaction network.
(Abelson, 1964; Kitts, 2006; Tsuji, 2002). Heterophobia states that individuals do not like other agents with a sufficiently dissimilar opinion (Byrne et al., 1986; Chen and Kenrick, 2002; Pilkington and Lydon, 1997; Rosenbaum, 1986).

However, there is a lack of microfoundation with respect to the implementation of the four concepts in the models. In particular, an analysis of the underlying mechanisms could be helpful with respect to modelling the interplay between the counterparts social influence/rejection and homophily/heterophobia. And even if only social influence and homophily are considered, Hegselmann and Krause (2005) show that different implementations of social influence (i.e. different ways of averaging) may significantly influence the dynamics. In this chapter, we close this gap by providing a microfoundation of social influence (and rejection) in opinion dynamics models based on cognitive dissonance.

2.2 A Microfoundation of Social Influence based on Cognitive Dissonance

According to the concept of social influence, an agent adjusts her opinion to the opinions of her interaction partners (the in-group). Models of opinion dynamics usually propose rules of how the agents’ opinions are affected by interaction with other agents. Thereby, social influence is directly implemented on the macro level. In contrast, we provide a microfoundation of social influence based on the social psychological concept of cognitive dissonance (Festinger, 1957). According to it, an individual strives towards consistency (consonance) in her cognitions which can be opinions, beliefs or knowledge. An example for inconsistent (dissonant) cognitions is to smoke while being aware of the health risk incurred. Festinger’s basic hypotheses are that “The existence of dissonance, being psychologically uncomfortable, will motivate the person to try to reduce the dissonance and achieve consonance.” and that “When dissonance is present, in addition to trying to reduce it, the person will actively avoid situations and information which would likely increase the dissonance.” (Festinger, 1957, p. 3).

With respect to opinion dynamics, we assume that each agent strives to have “similar” opinions compared to her in-group members. Consequently, different (relevant) opinions lead to cognitive dissonance. Further, “the strength of the pressures to reduce the dissonance is a function of the magnitude of the dissonance” (Festinger, 1957, p. 18). In the following model, we assume that the magnitude of dissonance for an individual resulting from a different opinion of an in-group member increases with the “distance” of the opinions and the intensity of the relation between the two individuals. In order to reduce the
### 2.2.1 The Model

Consider $n$ agents whereas agent $i$’s opinion is denoted by $x_i \in X$ with $i \in \{1, \ldots, n\}$. $X$ is the opinion space which may be finite or infinite. We require that the opinion space $X$ is metric, i.e. there exists a distance function $d : X \times X \rightarrow \mathbb{R}_+$ with the following standard properties:

\begin{align}
\text{(i)} & \quad d(x, y) = 0 \iff x = y, \\
\text{(ii)} & \quad d(x, y) = d(y, x), \\
\text{(iii)} & \quad d(x, y) \leq d(x, z) + d(z, y). \\
\end{align}

The first property reflects that a distance of zero between two opinions is only possible if the opinions are equal. Further, the order of the opinions does not influence the distance. The...
2.2. A Microfoundation of Social Influence based on Cognitive Dissonance

The third property reflects the so-called *triangle inequality*: The distance between two opinions is at least as high as the sum of the distances of each opinion to a third opinion. Note that one can define the discrete metric \( d_0 \) with 
\[
d_0(x, y) = 0 \text{ if } x = y \text{ and } d_0(x, y) = 1 \text{ otherwise}
\]
on an arbitrary opinion space. We refer to \( d(x, y) \) as the *opinion distance* between opinions \( x, y \in X \). In particular we allow that the opinions are vectors, i.e. \( X = (X_1, \ldots, X_m) \), where the components correspond to the opinions on \( m \) different subjects. If the opinions are real vectors \( (X \subset \mathbb{R}^m) \), we assume that \( d \) is the Euclidean distance if not stated otherwise, i.e.
\[
d(x, y) = \|x - y\|_2 = \sqrt{\sum_{1 \leq j \leq m} |x_j - y_j|^2}, \quad x = (x^{(1)}, \ldots, x^{(m)}), y = (y^{(1)}, \ldots, y^{(m)}).
\]

Further, we denote \( i \)'s in-group by \( I_i \subset I = \{1, \ldots, n\} \), and the agents in \( I_i \) are \( i \)'s *neighbors*. The in-groups induce a graph \( G \) on the set of agents: For agents \( i \) and \( j \), \((i, j)\) is in the set of edges \( E(G) \) of \( G \) if \( j \) is in \( i \)'s in-group, i.e. if \( j \in I_i \). We do not make any assumptions about how an agent’s in-group is constituted as this is not part of the process of social influence. It may be defined by a fixed external interaction network as in the voter model (Holley and Liggett, 1975), by the opinion profile according to bounded confidence (e.g. Hegselmann and Krause, 2002) or evolve over time according to a certain network formation process (e.g. Centola *et al.*, 2007; Vazquez *et al.*, 2008).

We assume that an opinion does not have an intrinsic value. Instead, the value of an opinion \( x \in X \) for agent \( i \) only depends on the cognitive dissonance caused by the opinion difference of \( x \) to opinions of the agents in \( i \)'s in-group. Considering agent \( i \) and her neighbor \( j \) with opinions \( x_i \) and \( x_j \), we assume that the magnitude of dissonance is a function of the opinion distance, i.e. it is captured by \( f^-(d(x_i, x_j)) \) whereas we refer to \( f^- : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) as the *distance dissonance function*. As we assume that the magnitude of dissonance strictly increases with the opinion distance, we require that \( f^- \) is strictly increasing. Further, the magnitude may depend on the intensity (or the weight) of the in-group relation between \( i \) and \( j \) which we represent by \( a_{ij} \in \mathbb{R}_+ \). According to Festinger (1957), we assume that the higher the intensity \( a_{ij} \) the higher the magnitude of dissonance caused by the opinion difference of \( i \) and \( j \). More precisely, it is \( a_{ij} f^-(d(x_i, x_j)) \). As an agent’s in-group may consist of more than one agent, the overall extent of dissonance \( u^-_i \) of \( i \) is the aggregation of the dissonances caused by opinion difference with respect to each of her neighbors. Here, we assume that
\[
u^-_i(x_1, \ldots, x_n) = \sum_{j \in I_i} a_{ij} f^-(d(x_i, x_j)), \quad (2.2)
\]
i.e. the dissonances are summed up. According to Festinger (1957), the agents strive to
minimize the cognitive dissonance $u_i^-$. In the following, we will analyze the structure of $u_i^-$ and investigate the resulting (myopic best response) dynamics.

### 2.2.2 Identifying Coordination as the Underlying Process

We can define an $n$-player game $G^-$ with strategy space $X$ and utility function $u_i = -u_i^-$ for each agent (player). As the maximization of utility $u_i$ is equivalent to the minimization of the cognitive dissonance $u_i^-$ for all players, we can investigate the process of social influence by analyzing $G^-$. First we consider a scenario with perfectly rational agents where the respective utility functions are common knowledge. In this context, we first neglect local effects by assuming that an agent’s utility of a certain opinion depends on all other agents’ opinions, i.e. $I_i = I \setminus \{i\}$. As $f^-$ is strictly increasing, a state where all agents have identical opinions is a strict Nash equilibrium of $G^-$. We denote such a state as consensus and define the consensus set as

$$C = \{(x, \ldots, x) \in X^n | x \in X\}. \quad (2.3)$$

Further, we denote the set of Nash equilibria of $G^-$ by $\Theta(G^-)$. As stated above, we have $C \subset \Theta(G^-)$. In general, $\Theta(G^-)$ not necessarily coincides with $C$:

**Example 2.1.** Consider four agents with opinion space $X = \mathbb{R}$ and opinions $x_1 = x_2 = 0$, $x_3 = x_4 = 1$ whereas the dissonance caused by deviation in opinion is proportional to the opinion distance, i.e. $f(z) = cz$ for all $z \in X$ with $c > 0$. If agent 1 perceives a higher dissonance for deviating from the opinion of agent 2 compared to the deviance to the opinions of agent 3 and agent 4 due to a more intense in-group relation with agent 2 (e.g. by $a_{12} = 1$, $a_{13} = a_{14} = 0.1$), the optimal decision for agent 1 given the other agents’ opinions is $x_1$. For analogous assumptions with respect to the other agents, $x_i$ is optimal for agent $i$ given the other agents’ opinion for $i = 1, \ldots, 4$, i.e. the non-consensus state $(x_1, \ldots, x_4)$ is a strict Nash equilibrium of $G^-$. However, we can provide sufficient conditions with respect to the distance dissonance function $f^-$ and the intensities $a_{ij}$ of in-group relations so that consensus states are the only Nash equilibria of $G^-$ in case of global interaction:

**Proposition 2.2.** If $I_i = I \setminus \{i\}$ and $a_{ij} = a$, then

$$\Theta(G^-) = C.$$
2.2. A Microfoundation of Social Influence based on Cognitive Dissonance

Proof. We only have to show \( \Theta(G^-) \subset C \). Without a loss of generality, we assume that \( a = 1 \) as we obtain

\[
u_i^{-}(x_1, \ldots, x_n) = \sum_{j \in I_i} f^{-}(d(x_i, x_j))
\]

for opinions \( x_1, \ldots, x_n \) in case the in-group influence is homogeneous. Hence, \( a > 0 \) does not affect the order of opinion profiles induced by cognitive dissonance.

Now consider a non-consensus opinion profile \( x = (x_1, \ldots, x_n) \). First suppose that we have

\[
u_{\text{max}}^{-} = \max_{j \in I} u_j^{-}(x) > \min_{j \in I} u_j^{-}(x) = u_{\text{min}}^{-}.
\]

Now choose an agent \( k \) with \( x_k = u_{\text{min}}^{-} \). With \( x' = (x_1, \ldots, x_n) \) whereas \( x'_i = x_k \) and \( x'_j = x_j \) for \( j \neq i \), we obtain

\[
u_i^{-}(x') = \sum_{j \in I \setminus \{i\}} f^{-}(d(x'_i, x'_j))
\]

\[
= \sum_{j \in I \setminus \{k\}} f^{-}(d(x'_k, x'_j))
\]

\[
\leq \sum_{j \in I \setminus \{k\}} f^{-}(d(x_k, x_j))
\]

\[
= u_k^{-}(x)
\]

\[
< u_i^{-}(x).
\]

In case there is \( u_{\text{max}}^{-} = u_{\text{min}}^{-} \), we choose agents \( i, k \) with \( x_i \neq x_k \) which is possible as \( x \) is not a consensus profile. With \( x' = (x_1, \ldots, x_n) \) whereas \( x'_i = x_k \) and \( x'_j = x_j \) for \( j \neq i \), we obtain

\[
u_i^{-}(x') = \sum_{j \in I \setminus \{i\}} f^{-}(d(x'_i, x'_j))
\]

\[
= \sum_{j \in I \setminus \{k\}} f^{-}(d(x'_k, x'_j))
\]

\[
< \sum_{j \in I \setminus \{k\}} f^{-}(d(x_k, x_j))
\]

\[
= u_k^{-}(x)
\]

\[
= u_i^{-}(x)
\]

where the inequality is strict as \( x_i \neq x_k \). Hence, \( x \) is not a Nash equilibrium as agent \( i \) can reduce her perceived dissonance for the given opinions of the other agents.

\[\square\]
Hence, all Nash equilibria are consensus states in case of homogeneous intensities of in-group relation. As shown in Example 2.1, this condition cannot be omitted.

Therefore, in case of homogeneous intensities of in-group relations, we can interpret $G^{-}$ as a pure coordination game in the sense that the Nash equilibria (which coincide with the consensus states) exhibit the same payoffs whereas the payoffs for all agents are also identical and maximal (no non-equilibrium state leads to a higher payoff for any agent). However, there may be focal points (Schelling, 1960), i.e. opinions that are in a sense more reasonable than others although the payoff in the correspondent consensus state is identical. If agent $i$ assumes that all other agents choose their respective opinion $X_j$ randomly according to a uniform distribution on the opinion space, she prefers an opinion $x_i^*$ that minimizes the expected dissonance, i.e.

$$x_i^* \in \text{argmin} \left\{ x \mid \sum_{j \in I_i} a_{ij} E(f^-(d(x, X_j))) \right\}. \quad (2.4)$$

For $2 \times 2$ coordination games, this corresponds to the concept of risk dominance of Harsanyi and Selten (1988). If we choose $X = [0, 1]$ according to many models of continuous opinion dynamics (cf. Hegselmann and Krause, 2002) and assume again global interaction, we obtain that for $x < 0.5$,

$$E\left(f^-(d(x, X_j))\right) = \int_0^1 f^-(d(x, y)) \, dy$$

$$= \int_0^{x+0.5} f^-(d(x, y)) \, dy + \int_{x+0.5}^1 f^-(d(x, y)) \, dy$$

$$= \int_{0.5-x}^1 f^-(d(0.5, y)) \, dy + \int_{x+0.5}^1 f^-(d(x, y)) \, dy$$

$$> \int_{0.5-x}^1 f^-(d(0.5, y)) \, dy + \int_0^{0.5-x} f^-(d(0.5, y)) \, dy$$

$$= E\left(f^-(d(0.5, X_j))\right).$$

Analogously, one can show that the same holds for $x > 0.5$. Hence, $x_i^*$ is the unique focal point in the sense of Eq. (2.4). As this holds for all agents, consensus with the average of the opinion space is the most reasonable equilibrium in case of global interaction with respect to minimizing expected dissonance if all agents assume that other agents choose their respective opinion randomly according to a uniform distribution. This holds independently of the distance dissonance function $f^-$. The reason is of course the simple structure of the opinion space. If we consider multidimensional opinions in a more complicated (although convex) opinion space, the average over all possible opinions is not necessarily a focal point with respect to Eq. (2.4). In case we can measure opinion distance only by the discrete metric as for example in the model of Axelrod (1997) (i.e. $d$ is
the discrete metric), all opinions are equivalent with respect to expected dissonance if the agents assume a uniform random distributions of opinions.

2.2.3 Myopic Best Response Dynamics

In the following, we reduce the degree of the agents’ rationality by assuming myopic best response dynamics: At time $t$, agent $i$ chooses an opinion $x_i(t + 1)$ that minimizes dissonance with respect to the opinion profile $x(t) = (x_1(t), \ldots, x_n(t))$ at time $t$ (without taking into account that other agents will do the same). Within this context of bounded rationality and evolving opinions, we assume that an agent does not solely take into account her neighbors’ opinions, but that she additionally perceives dissonance by deviating from her own opinion at time $t$. Again, this dissonance depends on the distance dissonance function $f^−$. It further increases with the agent’s confidence in her own opinion which is captured by $a_{ii}$. Thus, in contrast to the previous setting with perfectly rational agents, agent $i$ is contained in her in-group $I_i(t)$ which in general may also vary with time. Agent $i$’s perceived dissonance caused by opinion $y$ at time $t + 1$ with opinion profile $x(t)$ at time $t$ is therefore

$$u_i^−(y, x(t)) = \sum_{j \in I_i(t)} a_{ij} f^−(d(y, x_j)). \quad (2.5)$$

The opinion $x_i(t + 1)$ minimizes cognitive dissonance with respect to the opinion profile $x(t)$, i.e.

$$x_i(t + 1) \in \arg\max_{y \in X} u_i^−(y, x(t)). \quad (2.6)$$

We also refer to $x_i(t + 1)$ as a best response to the opinion profile $x(t)$. Note that the above best response dynamics exhibit a synchronous update of the agents’ opinions. Alternatively, we could choose one agent (randomly or in sequence) per timestep whose opinion is updated while the remaining agents’ opinions do not change. We refer to the latter case as asynchronous update of the agents’ opinions.

The basic assumption with respect to social influence is that interaction enhances similarity of the interaction partners. The following proposition shows that we can derive this assumption from the minimization of cognitive dissonance by the agents if the opinion space is convex.³

³A set is convex if the line between any two points of the set is contained in the set.
Proposition 2.3. Assume a convex opinion space \( X \subset \mathbb{R}^m \). Then for any agent \( i \), a best response \( x_i^* \) to an opinion profile \( x = (x_1, \ldots, x_n) \) is in the convex hull of \( x \), i.e.

\[
 x_i^* \in \left\{ y \in X \mid y = \sum_{j \in I_i} \lambda_j x_j, \; \lambda_j \geq 0, \; \sum_{j \in I_i} \lambda_j = 1 \right\}.
\]

Proof. Let \( y \in X \) denote an opinion that is not in the convex hull \( H \) of \( x \). As \( H \) is compact and \( d(\cdot, y) \) is continuous (\( d \) is the Euclidean metric), there is an opinion \( z \in H \) with (positive) minimum distance to \( y \), i.e. for all \( z' \in H \) we have

\[
 0 < d(z, y) \leq d(z', y). \tag{2.7}
\]

Now consider an agent \( j \). The opinion distance \( d(y, x_j) \) is greater than the distance \( d(z, x_j) \): Let \( y_p \) denote the orthogonal projection of \( y \) on the line \( l \) defined by \( x_j \) and \( z \), i.e. we have

\[
y_p = \lambda x_{j_1} + (1 - \lambda) z
\]

with \( \lambda \in \mathbb{R} \). If there was \( \lambda > 1 \), it would follow that \( d(x_j, y_p) < d(z, y_p) \) and therefore (by the Pythagorean theorem) \( d(x_j, y) < d(z, y) \) which is a contradiction to Eq. (2.7). If there was \( 0 < \lambda \leq 1 \), the Pythagorean theorem would yield that \( d(y_p, y) < d(z, y) \). As \( y_p \in H \) according to the convexity of \( H \), this is again a contradiction to Eq. (2.7). Hence, we obtain that \( \lambda \leq 0 \). Then, the Pythagorean theorem yields

\[
d(y, x_j)^2 = d(y, y_p)^2 + (d(x_j, z) + d(z, y_p))^2
\]

and therefore \( d(y, x_j) > d(z, x_j) \) as \( d(y, y_p) = d(z, y_p) = 0 \) would imply \( z = y \) which is a contradiction to \( y \not\in H \).

As \( f^- \) is strictly increasing in the opinion distance, this implies \( u^-_i(z, x) < u^-_i(y, x) \) so that \( y \) cannot be a best response to the opinion profile \( x \). \( \square \)

This proposition confirms the main hypothesis of social influence, namely that interaction increases similarity of the agents’ opinions. This holds independently of the choice of the distance dissonance function. In particular, it implies that the maximum distance will never increase during the process of best response dynamics. If the opinions are one dimensional (\( X \subset \mathbb{R} \)), the proposition implies that the best response \( x_i^* \) to an opinion profile \((x_1, \ldots, x_n)\) is a partial abstract mean according to the definition of Hegselmann and Krause (2005). Note that the proposition does not hold if the opinion space is not convex:

Example 2.4. Consider the discrete opinion space \( X = \{(0, 3), (6, 0), (6, 6), (7, 3)\} \), a population of \( n = 3 \) agents and assume that \( f^-(z) = z^2 \). Further, assume that the opinion
2.2. A Microfoundation of Social Influence based on Cognitive Dissonance

Figure 2.2: Visualization of Example 2.4: The dissonance minimizing opinion for agent 1 is $x_1'$ although it is not contained in the convex hull of the opinions $x_1$, $x_2$ and $x_3$.

profile at a time is $x = (x_1, x_2, x_3)$ with $x_1 = (0, 3), x_2 = (6, 0), x_3 = (6, 6)$ (see Figure 2.2). If the agents interact globally (i.e. $I_i = I$) and the intensity of the in-group relations is homogeneous (i.e. $a_{ij} = a$ and without a loss of generality$^4 a = 1$), we obtain that

$$u_1^-(0, 3, x) = (36 + 9) + (36 + 9) = 90$$
$$u_1^-(6, 0, x) = (36 + 9) + 36 = 81$$
$$u_1^-(6, 6, x) = (36 + 9) + 36 = 81$$
$$u_1^-(7, 3, x) = 49 + (1 + 9) + (1 + 9) = 69,$$

i.e. agent 1 perceives minimum dissonance if she chooses opinion $x_1' = (7, 3)$ which is not in the convex hull of the opinion profile $x$. Additionally, the maximum opinion distance in the new profile $(x_1', x_2, x_3)$ increased to 7 compared to $\sqrt{45} < 7$ in the old profile $x$.

Further, Proposition 2.3 guarantees the existence of a best response for each agent to any opinion profile if the distance dissonance function $f^-$ is continuous. However, this best response is not necessarily unique in general. Uniqueness can be attained if we require strict convexity for $f^-$:

**Proposition 2.5.** Assume a convex opinion space $X \subset \mathbb{R}^m$. If the distance dissonance function $f^-$ is continuous and strictly convex, there is always a unique best reply for each agent to a given opinion profile $x = (x_1, \ldots, x_n)$.

$^4$See the proof of Proposition 2.2.
Proof. As $u_i(\cdot, x)$ is continuous on the (compact) convex hull $H$ of $x$, there exists an optimal opinion $x_i^* \in H$ with

$$u_i(x_i^*, x) \geq u_i(y, x)$$

for all $y \in X$ according to Proposition 2.3. To guarantee the uniqueness of the optimal opinion, it is sufficient to show that $u_i(\cdot, x)$ is strictly convex for any opinion profile. Therefore, let $d_{x_j}(y)$ denote the distance of opinion $y \in X$ and the fixed opinion $x_j$. As $d_{x_j}$ is convex and $f$ is increasing and strictly convex, we obtain

$$f^-(d(\lambda y + (1 - \lambda) z, x_j)) \leq f^-(\lambda d(y, x_j) + (1 - \lambda) d(z, x_j))$$

$$< \lambda f^-(d(y, x_j)) + (1 - \lambda) f^-(d(z, x_j))$$

for all $j \neq i$, $y, z \in X$ and $\lambda \in [0, 1]$. Hence, the composition $f^o d$ is strictly convex. Therefore, $u_i(\cdot, x)$ is strictly convex as it is a positive linear combination of strictly convex functions.

Here, strict convexity of $f^-$ implies that the dissonance caused by opinion distance grows disproportionately: The marginal dissonance increases with the opinion distance. For example, if an agent chooses opinion 0.5 and has two neighbors with opinion zero and one neighbor with opinion one, the cognitive dissonance caused by opinion difference to the neighbor with opinion one is higher than the corresponding dissonance with respect to the agents with opinion zero if $f^-$ is strictly convex. Note that in case of a synchronous update, the myopic best response dynamics will immediately lead to consensus if $f^-$ is strictly convex and the interaction between the agents is global (i.e. if $I_i(t) = I$) and homogeneous (i.e. if $a_{ij} = a$) as the agents’ optimal opinions with respect to a given opinion profile are then identical.

So far, we did not specify a particular functional form of the distance dissonance function $f^-$. In the following, we determine the location of the dissonance minimizing opinion with respect to a given opinion profile in case $f^-$ is quadratic.

**Proposition 2.6.** Let $X \subset \mathbb{R}^m$ be a convex opinion space. Further, let the distance dissonance function be quadratic, i.e. $f^-(z) = z^2$. Then the best response to an opinion profile $x = (x_1, \ldots, x_n)$ for agent $i$ with in-group $I_i$ is

$$x^* = \frac{\sum_j a_{ij}x_j}{\sum_j a_{ij}}.$$

---

5The continuity of $u_i(\cdot, x)$ on the interior of $X$ is already guaranteed by the convexity of the distance dissonance function.
2.3. Relation to Models of Opinion Dynamics

Proof. Let \( x = (x_1, \ldots, x_n) \) denote an opinion profile with \( x_i = (x_{i1}, \ldots, x_{im}) \). For an agent \( i \) and \( y = (y_1, \ldots, y_m) \in X \) we define \( \tilde{u}_i^-(y) = u_i^-(y, x) \). According to our assumptions on \( f^- \), we have

\[
\tilde{u}_i^-(y) = \sum_{j \in I_i} a_{ij} f^- \left( d(y, x_j) \right)
\]

\[
= \sum_{j \in I_i} a_{ij} f^- \left( \sqrt{\sum_{1 \leq k \leq m} (y_k - x_{jk})^2} \right)
\]

\[
= \sum_{j \in I_i} a_{ij} \sum_{1 \leq k \leq m} (y_k - x_{jk})^2.
\]

To determine the utility maximizing opinion \( x^*_i = (x^*_{i1}, \ldots, x^*_{im}) \), we derive the first order conditions \( \left( \frac{\partial \tilde{u}_i^-}{\partial y} (x^*) = 0 \right) \) and obtain

\[
x^*_ik \sum_{j \in I_i} a_{ij} = \sum_{j \in I_i} a_{ij}x_{jk}, \quad 1 \leq k \leq m.
\]

Further, the Hessian of \( \tilde{u}_i \) is positive definite as

\[
\frac{\partial^2 \tilde{u}_i^-}{\partial y_k \partial y_l} = \begin{cases} 
2 \sum_{j \in I_i} a_{ij} & \text{if } k = l \\
0 & \text{if } k \neq l 
\end{cases}
\]

so that \( x^*_i \) is a local minimum of \( \tilde{u}_i^- \). As \( f^- \) is strictly convex, the same holds for \( \tilde{u}_i^- \) so that \( x^*_i \) is a unique global minimum of \( \tilde{u}_i^- \).

Hence, the best response to an opinion profile in case of a quadratic distance dissonance function is the weighted average of the current opinions whereas the weights correspond to the relative intensity of the corresponding in-group relation. This coincides with how social influence is usually incorporated in many continuous models of opinion dynamics (e.g. Deffuant et al., 2000; Hegselmann and Krause, 2002).

2.3 Relation to Models of Opinion Dynamics

We will show that the dynamics of the different models of opinion dynamics can be represented within the framework of the previous section as myopic best response dynamics with a particular distance dissonance function \( f^- \) and in-groups \( I_i \). Therefore, we provide a microfoundation for the implementation of social influence in these models which is based on cognitive dissonance.
2.3.1 Bounded Confidence Models

First, we consider the class of models of opinion formation that is known as bounded confidence models. In these models, agents only interact if their opinions are sufficiently similar. At each timestep, an agent adopts the (weighted) arithmetic mean of her opinion and the opinions of her interaction partners. The two most prominent models of bounded confidence are the Hegselmann-Krause model (Hegselmann and Krause, 2002; Krause, 2000) and the Deffuant-Weissbuch model (Deffuant et al., 2000; Weisbuch et al., 2002) which we refer to as the HK model and the DW model, respectively. Although the opinions can be multi-dimensional in both models, the opinion space is usually $X = [0, 1]$. The difference between the models is the update procedure of the agents’ opinions: In the HK model, agents update their opinions simultaneously at each timestep by adopting the arithmetic mean (by component) of their in-group members. Here, the in-group $I_{HK}^i(t)$ of agent $i$ at time $t$ depends on the current opinion profile $x(t)$: It contains those agents whose opinion distance to $i$’s opinion at time $t$ does not exceed a threshold $\varepsilon_i$, i.e.

$$I_{HK}^i(t) = \{ j \mid d(x_i(t), x_j(t)) < \varepsilon_i \}.$$

Often, one assumes homogeneous bounds of confidence meaning that all agents have identical thresholds ($\varepsilon_i = \varepsilon$). With $G_{HK}^t$ denoting the graph induced by the in-groups at time $t$, the new opinion profile at timestep $t+1$ can be formulated as

$$x_{HK}^{t+1} = A(G_{HK}^t) x(t)$$

whereas the components $A_{ij}(G)$ of the confidence matrix $A(G)$ for in-groups $I_1, \ldots, I_n$ and the corresponding graph $G$ are defined by

$$A_{ij}(G) = \begin{cases} \frac{1}{|I_i|} & \text{if } j \in I_i \\ 0 & \text{otherwise} \end{cases}.$$

The update mechanism of the DW model is similar in the sense that an agent’s new opinion is a weighted arithmetic mean of her in-group members’ opinions. However, only two randomly chosen distinct agents can interact (if their opinion distance does not exceed the threshold) at each timestep while all other agents keep their respective opinion. Hence, agent $i$’s in-group at time $t$ is defined as

$$I_{DW}^i(t) = \begin{cases} \{Z_1(t), Z_2(t)\} & \text{if } (i = Z_1(t) \text{ or } i = Z_2(t)) \text{ and } d(x_{Z_1(t)}, x_{Z_2(t)}) < \varepsilon_i \\ \{i\} & \text{otherwise} \end{cases}.$$

whereas $Z = (Z_1, Z_2)$ denotes a discrete time stochastic process with $Z(t)$ uniformly distributed on $I^2$.\(^6\) With respect to the average of the two opinions, the weight of an

\(^6\)Sometimes, the stochastic process is defined on $I^2 \setminus \{(i,i) \mid i \in I\}$, i.e. only distinct agent are chosen.
agent’s own opinion is usually specified by $1 - \mu$ with $\mu \in [0, 0.5]$ while the weight of the other agent’s opinion is $\mu$. Hence, the new opinion profile at timestep $t+1$ is

$$x_{DW}(t+1) = A(\mu)(G_{DW}(t)) \cdot x(t)$$

with $G_{HK}(t)$ denoting the graph induced by the in-groups $I_{i_{DW}}(t)$ and

$$A_{ij}(\mu)(G) = \begin{cases} 1 - \mu & \text{if } j = i \\ \frac{\mu}{|I_i| - 1} & \text{if } j \in I_i \setminus \{i\} \\ 0 & \text{otherwise} \end{cases}$$

denoting the bounded confidence matrix for in-groups $I_1, \ldots, I_n$ and the corresponding graph $G$. Note that for $\mu = 0.5$, the confidence matrix of a graph $G_{DW}$ according to the DW model equals the confidence matrix for unweighted averaging (i.e. $A_{i_{DW}}(0.5)(G) = A(G_{DW})$) as here the in-groups contain a maximum of two agents.

According to this representation, we can reformulate both models as a myopic best response dynamics with respect to the dissonance function

$$u_{i_{HK}}^{-1}(y, x(t)) = \sum_{j \in I_{i_{HK}}(t)} d(y, x_j)^2, \quad y \in [0, 1]^m$$

for the HK model and

$$u_{i_{DW}}^{-1}(y, x(t)) = \sum_{j \in I_{i_{DW}}(t)} A_{ij}(\mu)(G_{DW}(t)) d(y, x_j)^2, \quad y \in [0, 1]^m$$

for the DW model with $x(t)$ denoting the opinion profile at time $t$.

Proposition 2.6 guarantees that the best response profile to $x(t)$ is the weighted average corresponding to the weights of the quadratic distance dissonance functions.

The model of Axelrod (1997) also implements the bounded confidence mechanism although its opinion space and the distance function is discrete. In this model, each opinion is a list of $F$ features. For each feature, there are $Q$ traits which represent the alternative values the feature may have. Further, Axelrod assumes a fixed network of agents whereas the probability of interaction for two adjacent agents is proportional to their similarity measured by the proportion of identical features. In case two agents interact, one of the agents adopts a trait of a randomly selected feature of her interaction partner.

7Note that $A_{ij}(G_{HK}(t))$ is constant for all agents $j$ in agent $i$'s in-group within the HK model. Hence, this factor can be omitted in the utility function as strictly increasing transformations do not affect the induced preferences.

8Axelrod (1997) however uses the term “culture” instead of “opinion”.
To analyze this implementation of social influence from the perspective of our model of cognitive dissonance, we abstain from a formal derivation of the model dynamics as it is not unique with respect to Axelrod’s model. Instead, we sketch how the main properties of the implementation of social influence can be derived from our model. First, we can define the distance between two opinions as the number of different traits of the opinions. However, we can only determine whether two traits are equal or not: no further distinction of different traits of features is possible. Hence there is in general no unique best response to an opinion profile as the reduction of cognitive dissonance caused by the adoption of a (distinct) trait of the interaction partner does not depend on the feature the trait refers to. This is consistent with the random selection of the feature in Axelrod’s model. Further, the fact that only one feature is adopted can be explained by the assumption that deviation from her current opinion causes more cognitive dissonance than deviation from the opinions of other in-group members, i.e. $a_{ii} > a_{ij}$ with $j \neq i$. Finally, the distance dissonance function $f^-$ has to increase sufficiently slow as only one feature is adjusted in case of interaction independent of the number of different features.

### 2.3.2 Voter Models

A more simplified approach to the study of opinion dynamics can be found with voter models. Here, the opinion space usually consists of only two opinions. At each timestep, a randomly selected agent $i$ chooses a particular opinion $j$ with a probability $p = g(r_i^{(j)}(x))$ which depends on the local frequency $r_i^{(j)}$ of that opinion in $i$’s neighborhood $I_i$ for a given opinion profile $x$. Here, $g$ is called the response function. In case that $g$ is linear, we have a linear voter model. The linear voter model with $g(r_i(x)) = r_i^{(j)}(x)$ was introduced independently by Clifford and Sudbury (1973) and Holley and Liggett (1975). Examples of non-linear voter models can be found in Molofsky et al. (1999), Schweitzer and Behera (2009) and Stark et al. (2008a). Usually, the in-group structure does not change over time, i.e. there is in particular no influence of the network by the opinions which could be caused by homophily.\(^9\)

Let us assume that agent $i$’s cognitive dissonance $u_i$ caused by deviance with respect to a given opinion profile is defined according to Eq. (2.5). Further, we assume homogeneous intensity of in-group relations ($a_{ij} = a$). In case the agents are restricted to a deterministic choice of opinions, $i$’s best response to a given opinion profile $x = (x_1 \ldots, x_i)$ is to choose

---

\(^9\)An exception is the model of Vazquez et al. (2008).
2.3. Relation to Models of Opinion Dynamics

\[ x_i^* \]

\[ x_i^* = \begin{cases} 
0 & \text{if } r_i^{(0)} > r_i^{(1)} \\
1 & \text{if } r_i^{(1)} > r_i^{(0)}, 
\end{cases} \]

i.e. the opinion with maximum frequency in her in-group. If the local frequency of both opinions is identical, the best response is not unique and we have to specify a tie breaker rule or let the agent choose randomly.

In the following, instead of restricting the agents to a deterministic choice, we allow for mixed strategies, i.e. each agent \( i \) chooses a probability distribution on the opinion space. As there are only two opinions, a distribution can be represented by the vector \((p_i^{(0)}, p_i^{(1)})\) with \( p_i^{(x)} \) denoting the probability of choosing opinion \( x \in X \). As \( p_i^{(0)} + p_i^{(1)} = 1 \), we can define \( p_i = p_i^{(1)} \) whereas \( p_i^{(0)} = 1 - p_i \). However, only the final opinion is observed by the agents at the next timestep. If the agents minimize their expected dissonance according to their respective probability distribution, the best response is still to choose the opinion with maximum frequency in an agent’s in-group with probability one as

\[ E_p(u_i^-(Y, x)) = (1 - p) \sum_{j \in I_i \atop x_j = 0} f^-(1) + p \sum_{j \in I_i \atop x_j = 1} f^-(1) \]

with \( Y \) denoting the random choice which is distributed according to \( p \). In particular, the best response does not depend on the distance dissonance function \( f^- \). Instead of referring to expected dissonance, we assume that the agents minimize the cumulative dissonance caused by the expected deviances to their in-group members’ opinions according to the probability distribution \( p = p^{(1)} \):

\[ \tilde{u}_i^-(p, x) = \sum_{j \in I_i \atop x_j = 0} f^-(p d(1, x_j) + (1 - p) d(0, x_j)) = \sum_{j \in I_i \atop x_j = 0} f^-(p) + \sum_{j \in I_i \atop x_j = 1} f^-(1 - p). \]

Without affecting the induced preferences, we can normalize the cognitive dissonance perceived by an agent by the population size \( n \) and redefine

\[ \tilde{u}_i^-(p, x) = (1 - r_i(x)) f^-(p) + r_i(x) f^-(1 - p) \quad (2.8) \]

with \( r_i(x) \) denoting the proportion of agent \( i \)’s in-group members with opinion one in the opinion profile \( x \). In particular, the best response to a given opinion profile now depends crucially on the distance dissonance function \( f^- \). If \( f^- (z) = z^2 \), the optimal distribution \( p_i^* \) for agent \( i \) with respect to an opinion profile \( x \) has to satisfy the first order condition

\[(1 - r_i(x)) p_i^* = r_i(x) (1 - p_i^*). \]
Therefore, we obtain
\[ p_i^* = r_i(x), \]
that is the best response is to choose the probability distribution that corresponds to the local frequency of the opinions in an agent’s in-group. Thus, for a quadratic distance dissonance function \( f^- \), the myopic best response dynamics with respect to the utility function in Eq. (2.8) is identical to the linear voter model if at each timestep one agent is randomly selected to update her opinion.

In general, the shape of \( f^- \) analogously induces a voter model where the probability for an agent to select a particular opinion depends non-linearly on the local frequency of that opinion in the agent’s in-group. By Proposition 2.5 it follows that the optimal distribution with respect to any opinion profile is always unique if \( f^- \) is strictly convex.\(^{10}\) Further, we derive several properties of the best response to a given local frequency of the opinions in the following proposition:

**Proposition 2.7.** Let \( f^- \) be convex and let \( p_i^* = g(r_i(x)) \) denote an agent’s best response to given local opinion frequencies in the opinion profile \( x \). Then the following statements hold:

\[ (i) \quad g(1 - r_i(x)) = 1 - g(r_i(x)). \]
\[ (ii) \quad g(0) = 0, \quad g(0.5) = 0.5, \quad g(1) = 1. \]
\[ (iii) \quad g \text{ is increasing.} \]
\[ (iv) \quad g(p) \in (0, 1) \text{ for } p \in (0, 1). \]

**Proof.** (i) is an immediate consequence of the equivalence of both opinions and the fact that \( p_i^{(0)} + p_i^{(1)} = 1 \). Thus, we have in particular \( g(0.5) = 0.5 \). Further, as \( f^-(z) > 0 \) for \( z > 0 \) implies \( \tilde{u}_i(p, x) > 0 \) for \( p > 0 \), we obtain \( g(0) = 0 \). Using (i), this also leads to \( g(1) = 1 \). With respect to (iii), we denote \( p^* = g(r) \) with \( r \in [0, 1] \), i.e.

\[ (1 - r) f^-(p) + r f^-(1 - p) > (1 - r) f^-(p^*) + r f^-(1 - p^*) \]

for \( p \neq p^* \). For \( r' > r \) and \( p < p^* \) this implies

\[ \tilde{u}^-(p) - \tilde{u}^-(p^*) = (1 - r) f^-(p) + r f^-(1 - p) - (1 - r') f^-(p^*) - r' f^-(1 - p^*) \]
\[ > (r' - r) \left[ f^-(p) + f^-(1 - p^*) - f^-(1 - p) - f^-(p^*) \right] \]
\[ > 0 \]

\(^{10}\)Although the basic opinion space \( X = \{0, 1\} \) is not convex, we can fulfill the requirements of Proposition 2.5 by extending it to the interval \([0, 1]\). Here, we consider only best responses to “pure” opinion profiles \( (x_1, \ldots, x_n) \) with \( x_i \in \{0, 1\} \).
as $f^- \geq 0$ and $f^-$ is strictly increasing. Hence, $g$ is increasing. If there is no consensus in the in-group, i.e. if $r \in (0,1)$, the convexity of $f^-$ implies

$$f^-(p) = f^-(p \cdot 1 + (1-p) \cdot 0) \leq p f^-(1)$$

and therefore

$$\tilde{u}^-(1,x) = (1-r) f^-(1) > p f^-(1) \geq f^-(p) > \tilde{u}^-(p,x)$$

for $p \in (0,1-r)$. Thus, using (i), it follows that it is optimal to choose one opinion with probability one if and only if every in-group member exhibits that opinion.

Figure 2.3: Agent $i$’s optimal probability $g_\alpha(r_i)$ of choosing an opinion with respect to a local frequency $r_i$ of that opinion for $f^-(z) = z^\alpha$ and various values of $\alpha$. The myopic best response dynamics according to Eq. (2.8) correspond to (non-linear) voter models with response function $g_\alpha$.

In Figure 2.3 we depict the decision function $g$ induced by distance dissonance functions $f^-(z) = z^\alpha$ for various parameters $\alpha > 1$ which implies strict convexity of $f^-$. We observe that as $\alpha$ increases, a small difference in the local frequencies of the two opinions is less
Chapter 2. A Microfoundation of Social Influence

reflected in the respective opinions’ selection probabilities. The reason is that an increase in $\alpha$ leads to a decreased dissonance caused by small opinion differences. Hence, for $\alpha \to \infty$, the agents choose both opinions with equal probability unless there is consensus in their respective in-group. For small values of $\alpha$, we observe the opposite effect as then the dissonance caused by a small opinion difference is high. Thus, in the limit for $\alpha \to 1$, the agents choose the majority opinion with probability one.

If we only account for cognitive dissonance resulting from opinion difference to in-group members, the response function is always increasing according to Proposition 2.7. However, also voter models with decreasing or non-monotone response function are investigated (e.g. Molofsky et al., 1999; Schweitzer and Behera, 2009). Note that these types of non-linear voter models can be represented in our framework if we incorporate consonant perception (cf. Section 2.4).

An alternative way to deduce the dynamics of the linear voter model from the utility functions in Eq. (2.5) is to extend a result from Kosfeld (2002) in the context of a $2 \times 2$ coordination game without risk-dominant equilibrium. Here, following the framework of Rosenthal (1989) and the proportional imitation rule introduced by Schlag (1998), it was shown that the linear voter model corresponds to a stochastic process where the probabilities of switching between the two strategies are proportional to the payoff difference.\(^{11}\)

Similarly, we can show that an analogous mechanism with respect to the $n$-player coordination game with strategy spaces $S_i = \{0, 1\}$ and payoffs according to Eq. (2.5) leads to the dynamics of the linear voter model. Here, let $p_i^{(j)}(x)$, $j = 0, 1$, denote the probability that agent $i$ chooses opinion $j$ for a given opinion profile $x$. Following Rosenthal (1989), we define $u_i(y, x) = -u_i^-(y, x)$ for $y \in \{0, 1\}$ and require that

$$p_i^{(1)}(x) - p_i^{(0)}(x) = \lambda \left( u_i(1, x) - u_i(0, x) \right)$$

with $\lambda > 0$, i.e. the difference of the probabilities for the two opinions is proportional to the respective payoff difference. With respect to the latter term we obtain

$$(u_i(1, x) - u_i(0, x)) = \sum_{\substack{j \in I_i \mid x_j = 1 \forall j \neq i \mid x_j = 0}} f^-(1) - \sum_{\substack{j \in I_i \mid x_j = 1 \forall j \neq i \mid x_j = 0}} f^-(1) = (r_i^{(1)}(x) - r_i^{(0)}(x)) f^-(1).$$

Using Eq. (2.9) for $\lambda = f^-(1)^{-1}$ and $p_i^{(1)}(x) + p_i^{(0)}(x) = 1 = r_i^{(1)}(x) + r_i^{(0)}$, this leads to

$$p_i^{(1)}(x) = r_i^{(1)}(x).$$

---

\(^{11}\)If there is a risk dominant equilibrium, the process corresponds to a biased voter model (Schwartz, 1977).
i.e. each agent’s best response to a given opinion profile $x$ is to adopt an opinion with the probability that corresponds to that opinion’s relative frequency in her in-group according to $x$. Thus, if at each timestep an agent is selected to choose the optimal opinion distribution, this process coincides with the linear voter model.

### 2.4 Combining Consonant and Dissonant Perceptions

So far we explained how the concept of social influence can be derived from a model based on cognitive dissonance. Here, deviation from the opinion of in-group members leads to dissonance. However, an agent may perceive a high opinion distance as positive if she does not like the respective agent. Using the nomenclature of Festinger (1957), the opinion difference and the negative relationship are consonant. On the “macro-level”, interaction with disliked agents leads to rejection (Abelson, 1964; Kitts, 2006; Tsuji, 2002) which is also incorporated in models of opinion dynamics (Flache and Macy, 2006; Jager and Amblard, 2005; Kitts, 2006; Salzarulo, 2006). In the following, we integrate consonance in our model of cognitive dissonance so that the agents perceive dissonance due to opinion difference to in-group members and consonance in case of deviation from the opinions of agents they dislike (the out-group).

The extent of the consonance that agent $i$ perceives due to a certain opinion distance to out-group members is captured by the distance consonance function $f^+ : \mathbb{R}_+ \rightarrow \mathbb{R}_+$. Similar to the distance dissonance function we require that $f^+$ is strictly increasing in the opinion distance. Similar to our previous assumptions with respect to the agents’ respective in-group, we assume that the out-group agent $i$ denoted by $O_i$ is given and fixed. Again the intensity of an out-group relation between agents $i$ and $j$ is captured by $a_{ij} > 0$. Further, we sum up all consonance resulting from opinion differences with respect to out-group members, i.e. we define the overall consonance according to

$$u_i^+(x_1, \ldots, x_n) = \sum_{j \in O_i} a_{ij} f^+(d(x_i, x_j)).$$

Here we assume that $I_i$ and $O_i$ are disjoint, i.e. an agent cannot exhibit both an in-group and an out-group relation with another agent. Agent $i$’s utility of opinion $x_i$ for given opinions of the other agents is now a combination of the dissonance due to opinion differences with respect to her in-group members and the consonance due to opinion differences with respect to her out-group members:

$$u_i(x_1, \ldots, x_n) = u_i^+(x_1, \ldots, x_n) - u_i^-(x_1, \ldots, x_n).$$

(2.10)
Now we derive the myopic best response dynamics for quadratic distance consonance and dissonance functions. In doing so, we assume that the intensities of in- and out-group relations are homogeneous. Similar to Section 2.2.3, the agents perceive dissonance in case they change their opinion. The magnitude of dissonance depends on the parameter \( \alpha \in [0,1] \) which captures the agent’s weight of the dissonance resulting from the difference to her own opinion relatively to the consonance or dissonance caused by deviation from other agents’ opinions. More precisely, agent \( i \)’s utility of opinion \( y \in X = [0,1] \) for a given opinion profile \( x = (x_1, \ldots, x_n) \) is

\[
u_i(y, x) = -\alpha (y - x_i)^2 + (1 - \alpha) \left[ -\sum_{j \in I_i} (y - x_j)^2 + \sum_{j \in O_i} (y - x_j)^2 \right].
\] (2.11)

If the utility function is convex (i.e. its second derivative is non-negative), the optimal opinion with respect to any given opinion profile is either zero or one. This would imply that every agent’s choice is always restricted to one of these extreme values which seems inappropriate for further consideration as the spectrum of opinions would be reduced to only two possible values. Therefore we assume for any agent \( i \) that

\[
\alpha + (1 - \alpha) (|I_i| - |O_i|) > 0
\] (2.12)

to guarantee that her utility function is strictly concave. Hence, a sufficient condition for the strict concavity is that every agent’s in-group consists of at least as many members as her out-group. In case that this condition is not fulfilled, an agents weight \( \alpha \) of dissonance caused by deviance from her own opinion must be sufficiently high. With this assumption \( u_i \) has a global maximum given by

\[
x_{max} = \frac{\alpha x_i + (1 - \alpha) (\sum_{j \in I_i} x_j - \sum_{j \in O_i} x_j)}{\alpha + (1 - \alpha) (|I_i| - |O_i|)}.
\]

on \( \mathbb{R} \). As we restricted the opinion space to the interval \([0,1]\) agent \( i \)’s best response to the opinion profile \( x \) is

\[
x_i^* = \begin{cases} 
  x_{\text{max}} & \text{for } x_{\text{max}} \in [0,1] \\
  0 & \text{for } x_{\text{max}} < 0 \\
  1 & \text{for } x_{\text{max}} > 1 
\end{cases}
\]

Evidently, the myopic best response dynamics heavily depend on the in-/out-group structure of the agents which can be arbitrarily complex. In our simulations we generate random in- and out-groups considering the spatial relation between the agents which is defined by an additional network. In our setting, this network is a two dimensional lattice with periodic boundary conditions. The spatial effect is brought into play by assuming that an
agent is more likely to have (positive or negative) relations to other agents in a certain neighborhood compared to distant agents. Thus with \( N(i) \) denoting the neighborhood for agent \( i \), we define

\[
\begin{align*}
p_{i}^{+} &= P(j \in I(i) \mid j \in N(i)), \\
p_{i}^{-} &= P(j \in O(i) \mid j \in N(i)), \\
p_{2}^{+} &= P(j \in I(i) \mid j \notin N(i)), \\
p_{2}^{-} &= P(j \in O(i) \mid j \notin N(i)),
\end{align*}
\tag{2.13}
\]

as the probabilities for an agent \( j \neq i \) to be in the in- or out-group of agent \( i \). We further assume \( p_{1}^{+}, p_{1}^{-} > p_{2}^{+}, p_{2}^{-} \) which means that it is more likely for an agent to be in another’s in- or out-group if she is in that agent’s neighborhood (see Figure 2.4). Based on these probabilities, we construct a random in-/out-group structure for every agent.

Figure 2.4: Two agents are in each other’s in-group (out-group) with a probability \( p_{1}^{+} (p_{1}^{-}) \) if they are neighbors with respect to the Moore-Neighborhood of size 2. Otherwise the probability is \( p_{2}^{+} (p_{2}^{-}) \). Solid lines illustrate the central agent’s in-group relations while her out-group relations are represented by dashed lines. Note that \( p_{1}^{+}, p_{1}^{-} > p_{2}^{+}, p_{2}^{-} \), i.e. there are more relations to neighbors than to agents outside the neighborhood (source: Fent et al. (2007)).

In our simulations we always consider \( N = 900 \) agents, each centered in a Moore-neighborhood of size \( 13 \times 13 \). For \( \alpha = 0.9 \), Figure 2.5 shows the agents’ trajectories and the distribution of their opinions after the last simulation step with in-/out-group realizations for different values of the probabilities from equation (2.13). The first two examples show
Chapter 2. A Microfoundation of Social Influence

Figure 2.5: Trajectories and distributions of the agents’ respective opinion $x_i$ for different network parameters $p_1^+, p_1^-, p_2^+$ and $p_2^-$ after the last step of the simulation: (a,b): 50, (c,d): 500, (e,f): 2000 time steps. While we observe consensus in the first setting, an increase in the probability of non-neighborhood out-group relations leads to polarized opinions. If we reduce the probability of in-group relations in an agent’s neighborhood compared to the first setting, we observe a variety of opinions with peaks at the center and at the border of the spectrum of opinions (source: Fent et al. (2007)).
2.4. Combining Consonant and Dissonant Perceptions

Figure 2.6: Spatio-temporal evolution of the agents’ opinions $x_i$, which is coded in color scale. The probabilities for the in- and outgroup relations within and outside the neighborhood are chosen as $p_{11}^+ = 0.4$, $p_{1}^- = 0.2$, $p_{21}^+ = p_{2}^- = 0.05$, which allows for the spatial coexistence of a multitude of different opinions. The corresponding distribution is shown in Fig. 2.5.f. From this histogram, it is clear that there are at least three dominating peaks, which can be interpreted as different conventions (source: Fent et al. (2007)).
more simple cases: in Fig. 2.5.a and Fig. 2.5.b we observe that the agents very quickly converge to a consensus very close to 0.5. Here, the spectrum’s center is the opinion all agents finally conform to. This case always appears if the size of the agents’ in-group is large compared to that of their out-group size.

Fig. 2.5.c and Fig. 2.5.d show a setting where we find a wider spectrum of opinions. Here, every agent has about 5 times more friends than enemies in her neighborhood while outside of this the ratio is vice versa (with lower absolute numbers). So it is more likely for an agent to accord with her neighbors which also makes a greater likelihood of local clusters plausible. The simulations show that in this situation, we find about one third of the agents stabilizing at each end of the spectrum while the last third is almost equally distributed over [0, 1].

The most interesting situation is shown in Fig. 2.5.e and .f: the setting is similar to Figs. 2.5.a-b, only $p_1^+$ has been reduced from 0.6 to 0.4. This reduction of each agent’s in-group in her neighborhood reduces the attracting force between the agents sufficiently to avoid a global consensus. Instead of this, we find about three peaks in the distribution: one at the empirical mean of approximately 0.37 and two at the end of the interval whereas the peak at zero is higher than that at one.

For this situation, Fig. 2.6 depicts the spatial evolution of the agents’ opinions over time. Already after two iterations, the initial random distribution has been evened out to a level close to 0.5 for almost all agents - this also applies to most other observed settings. So we have a short timescale where attraction between the agents is dominant. On a second, larger timescale we observe a differentiation of the agents’ opinions: After 50 steps we find that agents in the upper region prefer values greater than 0.5 whereas at the opposite side, values lower than 0.5 are preferred. Furthermore, the empirical mean of the agents’ opinions decreases clearly under the initial value of 0.5. In the next picture we find the the upper left corner dominated by agents with values higher than 0.5 while the remaining area the majoritarian opinion is clearly lower than 0.5 and the overall empirical mean is already lower than 0.4. As we see in Fig. 2.6.e and .f, this situation remains stable for the rest of the simulation, so neither the contracting nor the dispersing forces prevail in driving the agents to total consensus or polarization respectively.\footnote{An animated computer simulation (in color) of the whole spatio-temporal evolution can be found at \url{http://www.sg.ethz.ch/research}.

Varying the parameter $\alpha$ (without violating Eq. (2.12)) to change the weight for an agent’s own opinion and that of her in- and out-group, respectively, did not change the results qualitatively and only affected the system’s time to reach its stationary state - a larger $\alpha$ increases an agent’s weight on her own opinion and extends this time.
2.5 Concluding Remarks

Although the computer simulations are not systematic, they cover interesting examples. As a major finding, they can explain the emergence and the stable spatio-temporal coexistence of different conventions (the peaks in the distribution of opinions) prevalent in certain subgroups of society. Further, the final distribution of opinions is smoother than in the opinion dynamics model of Jager and Amblard (2005) which also finds coexistence in a spatio-temporal setting caused by attractive and repulsive forces between the agents. We want to emphasize that the topology of the social network, in particular the in- and out-group structure, is crucial for the development of the agents’ opinions. For the model under consideration, it is possible that there is no convergence to a (quasi)stationary distribution, i.e. agents change their opinion constantly over time and the convention adjusts instantaneously. While this kind of scenario may also have some relevance, we argue that a convention should change only on time scales larger than the dynamics of the agents’ individual opinion (which is shown as a quasi-stationary phenomenon in this framework).

2.5 Concluding Remarks

In this chapter, we provided a microfoundation of the concept of social influence in a model based on cognitive dissonance. Here, our main assumption is that deviation from the opinions of in-group members leads to dissonance which the agents want to minimize. We show that interaction between agents can be interpreted as a coordination process and increases the similarity of their opinions under very general conditions. As social influence is an integral part in many models of opinion dynamics, we therefore contribute to their theoretical foundation. Further, we extend the model by also considering out-group relations where deviation from the opinions of out-group members is perceived as consonance by an agent. Therefore, interaction leads to an interplay of attractive and repulsive forces on the macro level. Depending on the network structure, this may lead to consensus, polarization or a broad, multimodal distribution of opinions. For example, an agent could either change her opinion or change her in-group structure depending on which mechanism leads to a higher reduction of cognitive dissonance.

However, our model is restricted to the effect of the agent network on the opinions while the network is external and does not change over time. Many models of opinion dynamics additionally assume that a change in the opinions feeds back to the agent network (see Figure 2.1). However, our model could be extended by allowing for a change of agent i’s in-group structure in order to reduce cognitive dissonance in addition to the possibility of changing her opinion. This change could either consist of a reduction of the in-group
relation’s intensity with respect to an agent $j$ with a different opinion (i.e. a decrease of $a_{ij}$) or the elimination of $j$ from $i$’s in-group.

Further, the analysis is very general and does not allow for a study of which special instance of the general model fosters or inhibits consensus – one of the main questions with respect to pure coordination norms. In the following chapter, we tailor an opinion dynamics model to the interaction between spatially concentrated firms based on our general framework. Further, we investigate whether a proposed mechanism of interaction increases the probability of a consensus which would be beneficial to all firms.
Chapter 3

An Application: Local Cultures in Clusters of Firms

Since the early 1990s, economic theory and policy has dedicated a lot of attention to the causes and effects of clusters, i.e. spatial concentrations of firms in one or a few related industries (Ellison and Glaeser, 1997). Driven by their impressive prosperity, the factors underlying cluster success were extensively studied in the hope of replicating areas like the Silicon Valley or London. Locating in a cluster provides several benefits for firms. However, the existence of these benefits hinges on a set of rules on acceptable business practice – the cluster’s ‘local culture’. For instance, there is a rule among banks located in Frankfurt that they should not hire personnel away from competitors. In the Silicon Valley, the local culture is such that firms openly exchange ideas even with direct competitors. In order to partake in the cluster and its dynamics, firms need to respect these rules, which has made it difficult for outside companies (e.g. multinational enterprises) to tap into the cluster through subsidiaries. While some work has been conducted to determine the nature and enforcement of such local cultures, their emergence remains far from understood. So far, anecdotal evidence suggests that it depends on agreement about the nature of desirable business practice.

To shed light on the emergence of local cultures, this chapter studies the emergence of the initial consensus. It argues that cluster firms have to interact with each other. As these interactions are not cost-free, they give rise to changes in firm behavior and different inter-firm networks that affect the future behavior of constituent firms. Both mechanisms can lead to convergent or divergent behavior. Those cases finding converging behavior among all or a majority of cluster firms constitute situations in which the basis for a local culture emerges. This is also important for the emergence of clustering benefits and therefore the economic viability of the cluster. To study the emergence of converging behavior (i.e. the
first step towards a local culture), we develop an opinion dynamics model with bounded confidence and group-influence as both aspects are required to mimic the dynamics of inter-firm interaction in clusters.

In doing so, we proceed as follows: Section 3.1 describes the case under study. The emergence and effect of local cultures is found to be similar to that of norms and conventions as both are means to solve co-ordination and cooperation problems. However, the costliness of inter-firm interactions requires some amendments to existing models (Section 3.2). In particular, it justifies the application of a bounded confidence model introduced in Section 3.3: Agents (= firms) will only interact if they are sufficiently close in their behavior as the risk of a costly transaction going wrong are too high otherwise. Moreover, agents are linked by ties stemming from previous interactions. Thanks to costly interaction, agents try to maintain these ‘in-groups’ and modify their behavior accordingly. Section 3.4 then presents the findings regarding the conditions for consensual behavior among all or a majority of agents. It is found that the existence and influence of in-groups fosters consensus, especially if agents would only interact with behaviorally similar actors. As a result, the nature of inter-firm interactions in clusters is conducive to consensus and thereby the emergence of a local culture.

### 3.1 The Need for Local Cultures

In economic geography, local cultures with rules like ”do not hire from competitors”, ”exchange ideas freely” or ”deliver only the highest quality” are a phenomenon characterizing clusters like Silicon Valley (IT), London (financial services) or Prato, Italy (textiles). The existence of clusters is tied to that of local cultures because the benefits to clustering are subject to various co-ordination and co-operation issues that are solved by the rules in the local culture.

Usually, cluster benefits relate to scale and specialization effects as well as positive externalities. The former (scale and specialization) emerge since companies in a cluster usually divide the production process. Rather than having all firms manufacture shoes, one specializes in soles while other provide laces, tops, linings and so on (Pyke et al., 1990). This division of labour leads to *scale and specialization* benefits as firms can achieve a greater output with a limited budget and become more efficient in their activities (Smith, 2003). Second, companies conduct competing and complementary activities under identical local conditions in the cluster. This leads to *positive externalities* in the diffusion of ideas and the availability of skilled labour.\(^1\)

\(^1\)Competitors experiment with different strategies under identical conditions. This allows for direct
3.1. The Need for Local Cultures

To generate these benefits, cluster firms have to overcome several dilemmas. For a division of labour to emerge, suppliers need a fair price and a sufficient market (Smith, 2003, p. 27). Moreover, the quantities provided by different firms have to be aligned. Positive externalities in knowledge and personnel are tied to respective investments in research and training. This is only viable if defection (free-riding) is limited. Akin to famous cooperation problems like the Kula Ring (Ziegler, 2007) or the Chicago diamond market (Coleman, 1990), the cluster literature argues, that these dilemmas are solved by rules on acceptable business practice that make up the local culture. While some work has determined what local cultures look like (e.g. Porter 1990; Pyke et al. 1990; Saxenian 1994) and how their enforcement can be ensured (Holländer, 1990; Kandel and Lazear, 1992) their emergence is far less understood.

It is usually suggested that firms learn about successful behavior when interacting with others. Successful past behavior is then repeated and possibly copied among firms in the cluster, which implies that gradually, this behavior spreads in the population. Once established, such a consensus on "good" behavior creates expectation about others' future behavior, thereby reducing frictions in interaction. Specific rules making up a local culture can finally emerge to foster and enforce this consensual behavior by monitoring and punishment of defecting agents (Maskell, 2001, p. 926). The first step towards a local culture is thus a behavior that is viewed as desirable enough to become the basis of rule-making. We argue that a good candidate for viable rules is a behavior already prevalent in the cluster, i.e. a behavior that is shared by (a majority of) local firms. As a result, consensus on a certain behavior constitutes the necessary condition for rule-making, rule enforcing and the emergence of local cultures. We investigate how a consensus on a specific behavior emerges in a cluster. In doing so, we argue that the division of labour in clusters requires interaction between firms to manufacture the product. As interactions are not cost free (Coase, 1937; Williamson, 1975), two mechanisms come into play.

First, interaction costs increase with the difference in firm behavior. Firms with similar behavior are likely to respond similarly to future developments. This makes it unnecessary to specify all possible scenarios in a contract thereby reducing the cost of interaction. If all interactions are equally beneficial, the increasing cost of interaction implies that (a) firms will not interact with all possible partners and (b) the partners to an interaction comparison of performance and selection of best practice. In addition, firms tackling the same or related problems may exchange knowledge through various mechanisms (Allen, 1983). Both aspects contribute to knowledge spillovers. Finally, many firms in a cluster increase the quality of the local labour pool by training activity and immigration of skilled people.

\footnote{Investment research and training requires an understanding that allows firms to capitalize on it, i.e. no exploitation of others' efforts or hiring away of personnel trained elsewhere.}
modify their behavior to become more similar. Second, costly interactions make it beneficial to maintain links with existing partners. Firms therefore become more embedded in networks emerging from their interaction history. Moreover, the desire to maintain these networks may constrain their behavior as firms seek to remain sufficiently similar to existing partners.

The consensus resulting from these mechanisms (if any) can take very different forms. It can reside with a behavior that is very co-operative, i.e. every firm strongly invests in activities subject to externalities and supplier-buyer relations are characterized by fairness. Such a situation provides high incentives to free-ride implying that this behavior is not self-enforcing. In other instances, consensus can reside with very defective behavior where all firms try to exploit one another as much as possible and do not investment in activities with externalities. This situation would provide no incentive for deviation (at least not to an individual firm), i.e. the local culture is self-enforcing.

3.2 Local Cultures as Norms or Conventions

Once established, local cultures fulfill the function of social norms or conventions insofar as they solve co-operation or co-ordination problems. As a result, their emergence mimics that of norms and conventions, which unfolds as follows: The first stage is build on consensus formation, where agents reach consensus on a certain behavior through different mechanisms like optimizing behavior (Opp, 1982; Weber, 1999), imitating or replicating successful strategies (Asch, 1956; Sherif, 1973) or through trial and error search (Demsetz, 1967). Once the consensual behavior spreads and remains in the population, it creates expectations about everyone’s future behavior, which reduce friction in interactions (Axelrod, 1986; Koford and Miller, 1991; Sugden, 1989). Depending on whether the behavior is self-enforcing (convention) or not (norm), the second stage of the process differs. For conventions, the emergence of consensus is sufficient. In case of norms, behavioral regularities have to result in a sense of “oughtness” (Opp, 2001a) that may eventually lead to an enforceable norm prescribing this behavior: “Thus, patterns of action emerge that then become normative [...]. Individuals comply with the new norm both for the original reason that the behavior was appealing, and also because it is now socially enforced” (Horne, 2001b, p. 6).

Depending on the nature of consensual behavior (self-enforcing or not), local cultures correspond to conventions or social norms. In either case, the first stage of their emergence process (consensus building) is identical. This makes models on the emergence of norms or convention applicable to our case. In the literature, most work on norms and con-
3.2. Local Cultures as Norms or Conventions

Conventions is based on game-theory with a smaller subset of research studying consensus formation (through voter models or bounded confidence approaches). The game-theoretic approach to norms and conventions (Fehr and Fischbacher, 2004; Holländer, 1990; Kandel and Lazear, 1992; Ullmann-Margalit, 1977; Voss, 2001) is not directly applicable to our case for several reasons. By focussing on the nature of the game (e.g. payoff structure, repetition) and underlying agent interaction, individual incentives and efficient outcomes are derived. Conventions correspond to situations where the optimal outcome is a non-unique Nash equilibrium. Depending on the underlying mechanisms, different equilibria may be selected (see e.g. Pujol et al., 2005; Shoham and Tennenholtz, 1992). Norms instead emerge in situations where the optimal solution is not an equilibrium outcome. The norm is viewed as the solution to the problems preventing a better outcome in that particular game.³

The focus on payoffs and incentives implies that there is a known payoff structure. In other words, agents know the consequences of their actions, anticipate the choices of others and are thereby able to determine, what kind of behavior leads to efficient outcomes. In the Prisoner’s Dilemma situation, cooperation is the ex-ante optimal behavior when jointly maximizing the players’ outcome. Moreover, game theory is less concerned with the emergence of a particular norm or convention but rather with its effect for the game’s equilibrium outcomes. As we cannot determine ex-ante payoff values for business strategy and our concern resides more with the emergence than the effect of local cultures, we focus on an approach that does not assume an ‘optimal’ behavior but where the value of an agent’s behavior only depends on the number of agents adhering to it and on that behavior’s compliance with the predominant behavior in the agent’s personal network. In this sense, any consensus among all agents is ‘good’ - regardless of the nature of consensual behavior. However, our approach can be linked to game theoretic models of coordination norms by our results in Chapter 2: As it is based on an opinion dynamics model that is an instance our general framework, the adjustment of the agents’ behavior is a coordination process in the sense that it can be represented as a best response to the actual behaviors with respect to the coordination game in Eq. (2.2).

There are several models studying the emergence of consensus from agent interaction (Axelrod, 1997; Deffuant et al., 2000; DeGroot, 1974; Hegselmann and Krause, 2002; Lehrer and Wagner, 1981). They fall into two main classes: Voter models and bounded confidence

³Candidate mechanisms are repeated games, reputation formation, signalling mechanisms or punishment (Nowak, 2006). To be effective, consensus about the behavior that ought to be adopted is needed (e.g. cooperation). To some, this would already constitute a norm (Opp, 2001a). In addition to strong information requirements, game-theoretic approaches thus need a general co-operation ‘norm’ to identify defectors.
models. In voter models, agents are characterized by a *discrete opinion* (a binary variable in most cases) and are embedded in a network of given topology. They may adopt other opinions according to their frequency in the agent’s neighborhood. In linear voter models, the transition towards a given opinion is directly proportional its local frequency. In non-linear voter models other types of frequency dependent behavior are possible (Schweitzer, 2007). While consensus is always reached in linear voter models, non-linear responses to the local frequency of an opinion may prevent (Schweitzer and Behera, 2009) or accelerate (Stark *et al.*, 2008a,b) consensus.\(^4\)

Another class of consensus models deals with *continuous opinions* \(x_i\) represented as a real number between 0 and 1 (Deffuant *et al.*, 2000). Two agents \(i\) and \(j\), randomly chosen at each timestep, can only interact if the difference in their opinion does not exceed a threshold value \(\varepsilon\). Rather than taking place on a predefined network, agent interactions are randomized and conditional here. This mechanism of ‘bounded confidence’ is applied by Hegselmann and Krause (2002) where all possible interactions take place simultaneously. As investigated by means of several approaches (e.g. Ben-Naim *et al.*, 2003; Lorenz, 2006) consensus then largely depends on the value of the key parameter \(\varepsilon\).\(^5\)

Our model, formalized in the following section, builds on the bounded confidence approach, but combines it with the consideration of the dynamics in an agent’s *social network*. This is based on a feedback mechanism between an agent’s behavior and her personal network: past interactions with partners from the agent’s in-group affect her individual behavior which in turn influences the structure of the in-group, iteratively. Hence, as the novel element, our model combines both *opinion dynamics* and *network dynamics* at the level of individual agents.

Both aspects relate to the fact that interaction between agents is not cost-free. First, we argued that costs increase with differences in agent behavior since many actions are beneficial as long as both agents behave in the same way.\(^6\) As a consequence, interactions are conditional on sufficient behavioral similarity. Moreover, interacting agents approach each other’s behavior to lower interaction cost. Second, costly interactions make keeping past partners very beneficial. As a result, agents will want to keep their past partners

\(^4\)Stark *et al.* (2008a,b) have discussed a modification of the linear voter model, where agents become more reluctant to change their current opinion the longer they have it. It was shown that this deceleration of the individual dynamics may, for certain growth rates of the reluctance, even accelerate the formation of consensus on the system’s level. This counterintuitive result is partly due to the agent’s *heterogeneity*, i.e. differences in their individual behavior.

\(^5\)Deffuant *et al.* (2005) incorporate an extension of the bounded confidence mechanism in a model of innovation diffusion. For a survey of results on bounded confidence models see Lorenz (2007a).

\(^6\)The greatest losses often result from diverging strategies (e.g. prisoner’s dilemma).
3.3. Modelling the Emergence of Consensual behavior

We argued before that costly interactions result in two effects, (i) consensus formation, i.e. optimizing behavior within a group to avoid friction, and (ii) network formation, i.e. optimizing the agents’ social network structure by deleting links with agents whose behavior largely deviates from one’s own, thus making interaction more costly (or creating links with agents whose behavior is more similar). The two interlinked dynamics are specified as follows.

3.3.1 Consensus Formation

In order to reflect the first consequence of costly interactions, we need a model that makes interaction conditional on agent behavior. Following Deffuant et al. (2000), agent i’s behavior $x_i$ is represented as a real number between 0 and 1. Thus, we are able to measure the distance between two agents' behavior and to model a gradual approach if these agents interact. This is different to the cultural dissemination framework of Axelrod (1997) where cultures constitute a finite, discrete and in general non-metric set. There, interaction between two agents can only lead to complete assimilation of behavior for one of them, whereas agents approach each other’s behavior in our model. We further define the behavior profile $x = (x_1, ..., x_n)$.

In our model, in accordance with Deffuant et al. (2000), two agents $i$ and $j$ are randomly chosen at each timestep. They can only interact if the difference in their behavior does not exceed a threshold value $\varepsilon$ which can be regarded as a measure of openness. Regarding agent interaction, we can also interpret $\varepsilon$ as the difference in behavior where the costs and benefits break even: With greater behavioral differences, interaction costs would increase while the benefits are assumed to be constant. Such an interaction would lead to a net cost to the agents involved and will therefore not occur.\footnote{As in Moscovici and Doise (1992) where opinions too far from the majority don’t enter group discussion.} As the benefits and costs are
identical for all agents, the necessary condition for an interaction of $i$ and $j$ becomes:

$$|x_i(t) - x_j(t)| < \varepsilon.$$  \hspace{1cm} (3.1)

If two agents interact, they try to maximize the benefits of this exchange. As benefits are constant and costs decrease with behavioral differences, the behavior of interacting agents becomes more similar as both approach each other by identical amounts:

$$x_i(t+1) = x_i(t) + \mu(x_j(t) - x_i(t))$$
$$x_j(t+1) = x_j(t) + \mu(x_i(t) - x_j(t)).$$  \hspace{1cm} (3.2)

The speed of approach in behavior depends on the parameter $\mu$. It reflects the well established phenomenon that interacting parties become more similar (e.g. Axelrod 1997; Macy and Skvoretz 1998; McPherson et al. 2001; Strang and Soule 1998).

The dynamics specified by Eqs. (3.1), (3.2) are referred to as the baseline model in the following, as they result in the known behavior already discussed by Deffuant et al. (2000). We now extend this model by introducing the second aspect: Costly interactions imply benefits to keeping past partners. This is modelled by aggregating each agent’s past interaction partners. Each agent $i$ thus has a set $I_i$ of other agents constituting her in-group, i.e. the agent’s acquainted partners. As the agent would like to interact with these partners later, she tries to keep her behavior sufficiently similar to them. As agent behavior changes with her interactions, we argue that the in-group exerts an influence on the agent’s future interactions. This is achieved by combining an agent’s behavior $x_i$ and the mean behavior of her in-group $\bar{x}_{I_i(t)}$ to determine the effective behavior

$$x_{i}^{\text{eff}}(t) = (1 - \alpha_i(t))x_i(t) + \alpha_i(t)\bar{x}_{I_i(t)}$$  \hspace{1cm} (3.3)

at time $t$. The use of the mean behavior is chosen to mirror that the agent is equally interested in interacting with any of her past partners.\footnote{This treatment of group influence by averaging has a long tradition. Formal models of group decision-making (French, 1956; Harary, 1959; Hegselmann and Krause, 2002; Lehrer, 1975; Lehrer and Wagner, 1981; Wagner, 1978) account for group influence by weighted averages. Social impact theory (Latané, 1981; Latané and Nowak, 1997) also constructs group influence by averaging. Similar to social impact theory, our model features a decreasing marginal group influence with respect to adding more agents to the in-group.}

In Eq. (3.3), $\alpha_i \in [0,1]$ corresponds to the influence of agent $i$’s in-group on her effective behavior. We use this parameter to mirror the strength of group influence. Based on the aforementioned notion that agents like to keep their past partners, the influence of the group would increase with its size. In this model, we define $\alpha_i$ endogenously by

$$\alpha_i(t) = \frac{|I_i(t)|}{|I_i(t)| + 1}. \hspace{1cm} (3.4)$$
Hence, we assume that each agent puts equal weight on her own and each in-group member’s behavior. If \( i \) has never interacted with another agent, her in-group is empty (\(|I_i| = 0\)) implying \( \alpha_i = 0 \). Hence, the effective behavior of agent \( i \) is then identical to her behavior \( (x_i^{\text{eff}} = x_i) \). Further, \( \alpha_i \) approaches 1 with growing in-group size \(|I_i|\). If the in-group is large, \( i \)'s effective behavior will therefore tend towards the average in-group behavior.

In our model, two agents wanting to interact now have to compare the distance between their effective behavior (influenced by their respective in-groups) instead of that of their own behavior. As a result, the necessary condition for interaction between agents becomes

\[
|x_i^{\text{eff}}(t) - x_j^{\text{eff}}(t)| < \varepsilon. \tag{3.5}
\]

If two agents with empty in-groups interact, this is identical to Eq. (3.1) as \( x_i^{\text{eff}} = x_i \) and \( x_j^{\text{eff}} = x_j \) for \( I_i = I_j = \emptyset \).

### 3.3.2 Network Formation

In addition to maintaining their in-group, agents also seek to expand it with suitable new partners. To specify this, we assume that in-groups are initially empty for all agents. Later, they evolve according to the agents’ interactions as follows: In each simulation step, two agents \( i \) and \( j \) are randomly selected. If Eq. (3.5) holds for them, they interact and are added to each other’s in-group (if they are not already contained). Over time, agents \( i \) and \( j \) may interact with different agents. Therefore, their effective behavior can be altered either directly due to a change of \( x_i \) and \( x_j \) resulting from interaction with other agents, or indirectly by interactions of agents in their in-group affecting the average behavior of the respective in-group. Thus, we may encounter a situation where agents \( i \) and \( j \) interacted at time \( t \) and were added to each other’s in-group while later at \( t' > t \), their effective behavior may be modified such that \( |x_i^{\text{eff}}(t') - x_j^{\text{eff}}(t')| \geq \varepsilon \). In this case, agents \( i \) and \( j \) could no longer interact when selected and would be removed from each other’s in-group.

Thus, if \( i \) and \( j \) are selected at time \( t \), we have

\[
I_i(t+1) = \begin{cases} 
I_i(t) \cup \{j\} & \text{if } |x_i^{\text{eff}}(t) - x_j^{\text{eff}}(t)| < \varepsilon \\
I_i(t) \setminus \{j\} & \text{if } |x_i^{\text{eff}}(t) - x_j^{\text{eff}}(t)| \geq \varepsilon 
\end{cases},
\]

\[
I_j(t+1) = \begin{cases} 
I_j(t) \cup \{i\} & \text{if } |x_i^{\text{eff}}(t) - x_j^{\text{eff}}(t)| < \varepsilon \\
I_j(t) \setminus \{i\} & \text{if } |x_i^{\text{eff}}(t) - x_j^{\text{eff}}(t)| \geq \varepsilon 
\end{cases}. \tag{3.6}
\]

Note that the in-group relation is symmetric but may not be transitive, i.e. agent \( i \) being
contained in agent \( j \)'s in-group and agent \( j \) being contained in agent \( k \)'s in-groups does not require \( k \) being in \( j \)'s in-group.

As indicated before, the behavior of interacting agents becomes more similar. This means that interaction at time \( t \) alters the agents' behavior according to Eq. (3.2). For \( t + 1 \), this also feeds back on the effective behavior of \( i \) and \( j \), Eq. (3.3), as well as on the effective behavior of agents whose in-group contains \( i \) or \( j \). The effect of modifying \( i \)'s behavior for her effective behavior will decrease with larger \( I_i \). Over time, the agent can interact with others in and outside her in-group if Eq. (3.5) is satisfied. This influences her behavior as well as the evolution of her in-group over time. Agents previously outside \( i \)'s in-group are added to the set \( I_i \) once \( i \) successfully interacts with them. The addition of new agents to \( I_i \) influences her effective behavior and thereby her potential for future interaction.\(^9\)

Thus this model provides a feedback mechanism between the agents' behavior and their in-group's structure.

In the context of local cultures, we are mainly interested in whether the dynamics lead to consensus or in quantifying the degree of heterogeneity in agent behavior. We therefore investigate how the results for the baseline model are affected by the two extensions proposed here, namely the evolving agent network and its feedback on agent behavior. As known, equilibrium outcomes depend on the key parameter \( \varepsilon \) which distinguishes between open-mindedness and narrow-mindedness of agents. Similar to the baseline model, high values of \( \varepsilon \) favor consensus or a small number of unrelated population subgroups (=components).\(^{10}\)

Note that in equilibrium, the dynamics always partition the agent population into a certain number of network components, where all agents within a component share the same behavior. Obviously, the difference between the behavior in any two components is at least the threshold \( \varepsilon \) as interaction between agents from different components would still be possible otherwise. Further, each agent’s in-group coincides with her respective network component: Agents within a component are fully connected but have no links to outside agents.

\(^9\)Many models in sociology build upon a reciprocal link of behavior and interaction. See Carley (1991); Coleman (1961, 1980); Friedkin and Johnsen (1990); Marsden and Friedkin (1993); Nowak et al. (1990).

\(^{10}\)For both models, this does not hold in general as the probability of consensus and the average maximum component size as a function of \( \varepsilon \) are not monotonically increasing (see Ben-Naim et al. (2003) for the baseline model).
3.4 Findings

As explained before, agents need to agree on a desirable business behavior to allow for a local culture and cluster benefits to emerge. This consensus results from past interactions and the successful behavior therein. The strength of any local culture thus relates to the spread of a specific behavior in the population. Therefore one quantity measured in our simulations is the frequency of consensus among all agents, i.e. how often all agents exhibit the same behavior in equilibrium. As the behavior profile does not converge within a finite number of timesteps, we call a behavior profile \( x \) consensus profile if the maximum distance between two agents’ respective behaviors is at most \( \varepsilon \). In this case, all agents’ behavior will finally converge to the mean behavior in the population. Similarly, we can define a sufficient condition for non-convergence. First, there must be two components whose distance is at least \( \varepsilon \), i.e. if there exist agents \( i \) and \( j \) with \( |x_i - x_j| \geq \varepsilon \) and there is no other agent whose behavior is between \( x_i \) and \( x_j \). Second, we require that there are no links between these \( \varepsilon \)-separated components. In this case, the respective limit behaviors for \( i \) and \( j \) even for large \( t \) cannot coincide. However, we should also take into account to what extent there is a shared behavior among the agents in case of no consensus. We measure this by the size of the largest component of agents with identical behavior. For example, a situation where 95% of the agents share the same behavior is much closer to consensus than one where the population of agents splits into three equally sized components with different behavior.

We analyze the local cultures model with respect to these quantities by means of computer simulations and compare the results to the baseline model that has no feedback between behavior and network. The key parameter varied is \( \varepsilon \), high values of which characterize the openmindedness of the agents.

The main result relates to the question whether the feedback mechanism between the agents’ behavior and their network structure (group influence) is beneficial in the sense that a common behavior is fostered. Our simulations (see Figure 3.1 and Figure 3.2) show that for more narrow-minded agents, i.e. small thresholds \( \varepsilon \), the group influence results in both a higher average frequency of consensus and a larger maximum component size as compared to the baseline model. Hence, the mechanism introduced in our model increases the likelihood of consensus formation.

This result, interestingly and counterintuitively, changes for larger thresholds and therefore more open-minded agents. Here, the feedback mechanism weakens the emergence of a local culture in general. Agents subject to group influence reach less consensus on average than in the baseline model. We note, however, that the mechanism’s effect differs between the
Figure 3.1: Average frequency of consensus dependent on $\epsilon$ for 5000 runs and different values of $n$ and $\mu$. The agents’ initial behavior is random according to a uniform distribution, the initial network is empty. For narrow-minded agents (small $\epsilon$), the group influence fosters consensus while for open-minded agents (high $\epsilon$), the probability of identical behavior is lower than in the baseline model without group influence (source: Groeber et al. (2009)).
Figure 3.2: Average maximum relative component size dependent on $\epsilon$ for 5000 runs and different values of $n$ and $\mu$. The agents’ initial behavior is random according to a uniform distribution, the initial network is empty. For narrow-minded agents (small $\epsilon$), our model increases the average maximum component size compared to the baseline model without group influence. For open-minded agents (high $\epsilon$), there is almost no difference between the two models with respect to the average maximum component size (source: Groeber et al. (2009)).
Figure 3.3: Standard deviation of the maximum relative group size dependent on $\epsilon$ for 5000 runs and different values of $n$ and $\mu$. The agents’ initial behavior is random according to a uniform distribution, the initial network is empty (source: Groeber et al. (2009)).

two measures: While the frequency of consensus is significantly decreased (Figure 3.1), the effect on the maximum component size is much smaller (Figure 3.2). To explain the influence of agents open-mindedness, we argue that the feedback mechanism in the local cultures model implies two opposed effects compared to the baseline model. On the one
hand, agents with “extreme” initial behavior (i.e. an initial behavior close to zero or one) are less likely to interact with other agents and are therefore more likely to stay in that border area. The longer these agents remain in isolation (i.e. without interacting), the denser the network of other agents becomes, implying more averaging of behavior in determining the effective behavior of these networked agents. As more averaging leads to values closer to the mean, there are fewer and fewer agents within the interaction range of any isolated agent (as compared to the baseline model), and full consensus becomes more unlikely.

On the other hand, the feedback mechanism fosters consensus by increasing the coalescence of subpopulations with different behavior, i.e. components within the network. To illustrate this, consider the simulation depicted in Figure 3.4.\textsuperscript{11} In Figure 3.4(a), there are two nearly separated components in our model, the upper, smaller one with a higher average behavior, the lower, larger one with a lower average behavior. The two links that connect these components would not persist in the baseline model as the respective nodes’ difference in terms of their own behavior is above the threshold. However, as this is not the case for their effective behavior, the involved agents can still interact in our model. Hence, the two agents from the upper component still influence agents in the lower component by increasing their effective behavior (compared to their other neighbors whose behavior is lower). For the same reason, the two upper agents’ effective behavior is decreased by its neighbors from the lower component. Thus, they could establish further connections to the lower component. Nevertheless, interaction with agents from the upper component would increase their behavior and hence increase the distance to their neighbors from the lower component.\textsuperscript{12} Therefore, whether the two components stay connected and finally evolve to a complete graph or become separated depends on which nodes are chosen in the near future, i.e. is a path dependent process. Any interaction between the different components increases the probability of their coalescence, any interaction within the same component decreases that probability. In our example, one agent can establish further connections to the lower component (Figure 3.4(b)) and in return enables its neighbors from different components to interact (Figure 3.4(c)). Very quickly, more and more agents from the different components interact, become more similar and finally make the components coalesce (Figure 3.4(d)). This effect of coalescing components is also apparent in a higher variance of the maximum component size for narrow-minded agents as compared

\textsuperscript{11}See \url{http://intern.sg.ethz.ch/publications/local_cultures/web-cultures.html} for a video of this simulation. For a dynamic network layout, we used the \textit{arf} algorithm (Geipel, 2007).

\textsuperscript{12}This increase would have two reasons: first the increase of their own behavior, second the decrease of their lower component’s neighbors’ influence on their effective behavior as the share of lower component agents of the neighborhood also decreases.
Chapter 3. An Application: Local Cultures in Clusters of Firms

Figure 3.4: Network evolution for a simulation with 50 agents and $\varepsilon = 0.3$ at different timesteps. The agents’ initial behavior is random according to a uniform distribution, the initial network is empty. A node’s colour indicates the respective agent’s behavior (white=0, black=1). A green dashed link denotes that the respective agents’ difference in effective behavior is below the threshold while their respective own behaviors differ more than $\varepsilon$. Thus, such a link persists in the local cultures model but would be deleted in the baseline model. A red dotted link indicates that the respective agents’ difference in effective behavior is above the threshold, i.e. the link would be deleted if the respective agents were chosen at that timestep (source: Groeber et al. (2009)).
to the baseline model (cf. Figure 3.3).

Which of these effects decides over the strength of a local culture depends on the threshold $\varepsilon$: For narrow-minded agents, the baseline model is generally more likely to obtain several components instead of consensus. Thus, in the local cultures model, the increased probability of the coalescence of components increases both the frequency of consensus and the maximum component size. For open-minded agents, this effect vanishes because of the greater ex-ante likelihood of consensus in the baseline model. In this situation, the effect of isolation of agents with extreme behavior comes into play: While both models favor consensus in general, it is more likely for the local cultures model to find agents at the spectrum’s borders being separated from the other agents because of the faster dynamics towards the center. Therefore, the frequency of consensus is lower in this model. On the other hand there are only few agents separated from the majority, so the maximum component size is only slightly decreased by the feedback mechanism in the local cultures model. Hence, if we only consider this quantity to measure a local culture’s magnitude, the feedback mechanism significantly strengthens a local culture for narrow-minded agents and only slightly weakens it for open-minded agents.

What is the effect of variations to the population size $n$ and the convergence speed $\mu$ on our findings? To explain this we consider how both parameters affect the consensus frequency in our model and the baseline model. In both cases, an increase in the population size usually leads to a decreased consensus probability as it becomes more likely that a single agent with extreme initial behavior is separated from the rest of the population. For open-minded agents, this effect is amplified by the the faster dynamics towards the center in the local cultures model. Thus, we observe that the reduction in frequency of consensus is greater than in the baseline model (cf. Figure 3.1). With respect to the average maximum component size, an increase in the population size increases the advantage of the local cultures model compared to the baseline model for narrow-minded agents (cf. Figure 3.2). In this case, the increased number of agents leads to a higher probability for bridging links between two almost separated components. Hence, these components more often coalesce and thereby increase the difference between the maximum component size in our model and that in the baseline model as $n$ grows. This is also indicated by the increased distance of the two models’ respective variance peak (cf. Figure 3.3). If we decrease the convergence speed $\mu$, we observe an increase in the consensus frequency for both models for all thresholds $\varepsilon$ as the agents’ behavior moves slower towards the center. With respect to the maximum component size, this only holds for the baseline model. Figure 3.5 shows that this quantity decreases for narrow-minded agents, i.e. if $\varepsilon$ is small. The reason is that the coalescence of components becomes less likely for smaller values of $\mu$ as interaction with a bridge link between two almost separated components becomes less effective in this
Chapter 3. An Application: Local Cultures in Clusters of Firms

Figure 3.5: Average maximum relative component size dependent on $\epsilon$ for 5000 runs, 500 agents and different values of $\mu$. The agents’ initial behavior is random according to a uniform distribution, the initial network is empty. For the baseline model without group influence, a decrease of the convergence speed $\mu$ leads to an increase of the maximum component size for all $\epsilon$. This holds only for open-minded agents (high $\epsilon$) in case of our model (source: Groeber et al. (2009)).

3.5 Concluding Remarks

This chapter set out to study the emergence of local cultures. To do so, it focused on the first stage of the process where agents need to obtain consensus on acceptable business practice. Within a bounded-confidence model of opinion dynamics, we added a feedback mechanism between a agent’s behavior and the evolving agent network. The effect of this mechanism depends on the value of the interaction threshold $\epsilon$. In comparison with the baseline model (Deffuant et al., 2000), our feedback increased the likelihood of consensus for narrow minded agents (small $\epsilon$) as the group effect may foster a coalescence of otherwise separated components. For open-minded agents (large $\epsilon$), the likelihood of consensus decreased because the group effect worked to speed up convergence as compared to the baseline case. In some instances, this convergence was too fast for all agents to reach consensus. However, this constellation still had a substantial proportion of component agents finding consensus (the maximum component size was almost as large as in the baseline case). The fact that all component agents want to maintain their networks thus leads to behavioral constraints that may impede full consensus for very open-minded agents but increases it for narrow-minded ones. As local cultures can also emerge within a
sub-population, the aforementioned results suggest that the desire to maintain interaction networks has a positive effect on the emergence of (full or partial) consensus, which would then form the basis of a local culture.

With respect to model extensions, two aspects spring to mind. First, one could investigate the effect of heterogeneity among agents regarding their open-mindedness ($\varepsilon$). Recent contributions (Lorenz, 2008) suggest that heterogeneity plays a substantial role for the likelihood of consensus in the baseline model. Second, one could introduce non-empty initial networks (in-groups) to proxy that entrepreneurs in components often have an initial set of acquaintances from living in the area or studying in the same university. Given the contrasting effect of groups on consensus, it would be interesting to investigate in how far non-empty initial in-groups affect it. In a more general theoretic context, it would pay to apply this model to the emergence of norms and conventions in general. Given the more expansive body of research in this field, opportunities for benchmarking the model’s results against other existing studies will probably arise. It would be particularly interesting to see whether our counterintuitive result on group-effects and consensus is applicable to other constellations.

In the following chapter, the focus is shifted from coordination norm to cooperation norms. Here, the key problem is longer the selection of one of many equivalent equilibria as the optimal state is usually unique. However, it is not a Nash equilibrium and therefore not self-enforcing: There is an incentive for a selfish individual to deviate from the optimal state. Hence, the enforcement of cooperation norms is a key issue. In the second part of the thesis, we investigate the effect of subjective beliefs about the extent of norm violation on normative behavior when norm violation is not directly observable and can only be detected by an external inspector.
Chapter 4

Cooperation Norms and the Dark Figure of Norm Violation

Recently, there has been progress in understanding the underlying mechanisms of social order. Laboratory experiments revealed that altruistic punishment can solve the second order collective good problem of norm enforcement Fehr and Gächter (2002). However, the detection problem with respect to norm breakers induces a dark figure of norm violation which has been overlooked in most of these recent advances. Considerations about the unknown field of undetected deviance are not new to sociology: Examples are the debate between strain theory (Merton, 1957) and labeling theory (Becker, 1963; Kitsuse, 1962; Lemert, 1967; Tittle, 1980, 1969) or Popitz (1968) who investigates how beliefs of the extent of unknown norm violations influence normative behavior. However, a precise theoretical and formal description of the dark figure of norm violation on normative behavior is lacking. Further, labeling theory and Popitz (1968) constitute an isolated view of inspectors that shall detect norm violation and norm targets respectively. The game theoretic approach by Tsebelis (1990) investigates the interaction between norm targets and inspectors. Nevertheless, this approach neglects the detection problem as it is assumed that detection is guaranteed in case of inspection.

In this chapter, we present a mathematical model to analyze the interaction between subjective beliefs about the dark field on one hand and delinquency and control on the other. Figure 4.1 conceptualizes the relationship between labeling theory, Popitz (1968), Tsebelis (1990) and our approach. We can derive an inverse self fulfilling prophecy effect: if inspectors believe in a large unknown extent of crime, there will be less crime. Likewise, if norm breakers believe in a large unknown extent, the actual crime rate will be reduced in most cases. Furthermore, we can show that the dynamics between crime and control can stabilize at very different levels. These findings can explain why social norms are
Figure 4.1: Conceptualization of the relationship between Popitz (1968), labeling theory, Tsebelis (1990) and our approach. Popitz (1968) (dashed arrows) focuses on the side of norm targets: A higher expectation of the extent of norm violation (less “moral confidence”) increases violation of a social norm itself. Labeling theory (dotted arrows) focuses on the side of inspectors: Their confidence in the normative behavior of norm targets (or a subgroup of norm targets) influences the control behavior and therewith the rate of detected norm violations in the population (or in a subgroup of norm targets). Tsebelis (1990) (solid arrow) considers interaction between norm targets and inspectors, but neglects the detection problem and therefore a dark figure of norm violation. Our model brings all three approaches together and enables the analysis of their interaction.

adhered and control is performed meticulously in some societies, while there is disorder, delinquency and little control in others.

4.1 A Public Goods Game with Unobservable Actions

In our model, the population consists of \( n \) norm targets (agents) and one inspection institution, called the the inspector. Although we consider one singular inspection institution, it may consist of several inspectors. All inspections actions refer to the whole corporate actor, i.e. the inspection institution. We assume that at each time step \( t \), a norm target’s
strategy \( x_i(t) = (d_i(t), e_{a_i}(t)) \) consists of two components: First there is a binary choice between violating the norm \( (d_i = 1) \), and adhering to the norm \( (d_i = 0) \). We refer to \( d_i \) as the behavior of norm target \( i \). Second, \( e_{a_i} \in [0, \infty) \) captures the norm target’s expenditures on concealing her fraud. Here, a rational individual’s choice of adhering to the norm implies no expenditures on concealing the norm violation, i.e. \( e_{a_i} = 0 \) if \( d_i = 0 \). The inspector has no option to refrain from inspection. Hence, her strategy is restricted to her expenditures on control effort \( e_c(t) \in [0, \infty) \). An increase of control effort could reflect a larger sample of inspected norm targets or more sophisticated controls.

4.1.1 Detection Probability as a Result of Interaction

In our model we assume that detection of norm violation is neither guaranteed nor occurs with a fixed probability. Instead, the detection probability depends on the interaction between the norm targets and the inspector: For a norm violator \( i \) \( (d_i(t) = 1) \), the probability of being detected by the inspector at time \( t \) is

\[
p_i(e_{a_i}(t), e_c(t), q(t))
\]

where

\[
q(t) = \frac{1}{n} \sum_i d_i(t)
\]

denotes the proportion of norm violators among the norm targets. Hence, the detection probability depends on the strategy of both parties: It varies with the norm target’s concealment effort and the inspector’s detection effort. In addition, the detection probability may also depend on the overall strength of the norm captured by how many norm targets adhere to it, for example in the case of indirect detections. By this we mean that in addition to a direct detection of norm target \( i \)’s norm violation, it is also possible to discover indirectly the norm violation by the detection of a distinct norm target \( j \). The reason is that \( i \) and \( j \) might have a common supplier of instruments for norm violation or concealment, e.g. a common provider of drugs in case of doping or respective accounts for illegal earnings at a common bank. An example are the occurrences in the run-up to the Tour de France in 2006 where many cyclists were suspended after a physician was accused by the police of conducting autologous blood transfusions.

This internalization of the detection probability of norm violation is different from the mechanisms in present models of crime, tax evasion or doping. Considering the inspection game of Tsebelis (1990), the inspector can determine this probability while it is not
directly influenced by the norm targets.\footnote{\textit{1}} In the case of tax evasion, an endogenous audit probability\footnote{\textit{2}} was already discussed in the standard model (Allingham and Sandmo, 1972). Later models, for a survey see Andreoni \textit{et al.} (1998), enable the tax authority to choose an audit probability depending on the amount of declared income by a norm target. However, although cheating taxpayers often undertake substantial effort to conceal their evasion and therefore decrease the probability that their norm violation is detected in an audition (Andreoni \textit{et al.}, 1998, p. 836), these models do not provide a mechanism that reflects a cheating taxpayer’s influence on the detection probability. Moreover, in the context of tax evasion, one could criticize that a norm target’s behavior $d_i$ is a binary variable. In most models, taxpayers can choose different extents of norm violation by reporting a certain amount of income to the tax authority. Hence, $d_i = 0$ corresponds to truthfully reporting the income whereas any value below the true income constitutes norm violation. Hence, the difference between real and reported income (the \textit{tax gap}) indicates the extent of fraud. Our assumptions imply that a norm violator is restricted to a fixed tax gap. Hence, she can only vary the quality of concealing her fraud in the sense of influencing the detection probability but not the extent of norm violation in our model. Nevertheless, some models (e.g. Beck and Jung, 1989; Cronshaw and Alm, 1995) assume that income can be either “high” or “low” and thus make assumptions similar to ours.

\subsection{4.1.2 Modeling the Norm Targets and the Inspector}

For norm target $i$, the expected utility of a strategy $x_i = (d_i, e_{a_i})$ is defined by

$$u_a(x_i) = d_i \left( b_a - sp(e_{a_i}, e_c, q) - e_{a_i} - c_i \right) - r b_a q$$

(4.3)

for fixed control effort $e_c$ and fixed proportion of norm violators $q$. Here, $b_a > 0$ denotes the benefit of norm violation if the norm target is not detected. Norm violation imposes a disadvantage on all norm targets numbered by $rb_aq$ whereas the factor $r \geq 1$ measures the extent of welfare loss for the population caused by norm violation and corresponds to the multiplier for public contributions in the standard public goods game.\footnote{\textit{3}} $s > 0$ is the punishment imposed on a norm violator when her fraud is detected.

\footnote{\textit{1}}More precisely, the inspector can choose the probability of examining a norm target. In case of inspection, norm violation is always detected. In a game theoretical setting with complete information, a representative agent chooses the probability of norm violation.

\footnote{\textit{2}}It is assumed that the tax authority can direct costly audits that guarantee the detection of norm violation whereas tax evasion remains undiscovered without audit.

\footnote{\textit{3}}One might claim that norm violation can also lead to a welfare increase for the norm targets (e.g. induced by $r < 1$). With respect to doping, extraordinary achievements enhanced by drug use might lead to an increase of the aggregated income of all athletes by additional advertising revenue.
Additional to the expenditures $e_a$ for the concealment effort, a norm violator $i$ receives a disutility $c_i > 0$. This disutility is randomly assigned to each norm target according to a distribution function $F$ and is therefore not explained by the model. $c_i$ reflects norm target $i$’s initial adherence to the norm and does not change over time. It captures the extent to which the benefit of norm violation has to exceed the expenditures on concealment and the expected sanction cost if norm violation is detected for norm target $i$. For example, some athletes know that doping would pay off but have internalized the anti-doping norm to an extent that makes them refrain from drug use. Similarly, there are taxpayers who do honestly report their income despite considerable economic incentives for norm violation (Andreoni et al., 1998, p. 822). Note that by these assumptions, an increase in opportunity costs for adhering to the norm (e.g. by reduced control effort leading to a decrease of detection probability for all norm targets) leads to an increase of norm violation in the population. Therefore, the norm targets’ behavior is in accordance with the low-cost hypothesis (cf. Diekmann and Preisendörfer, 2003; North, 1986) which postulates that the congruence of the acceptance of a social norm with the corresponding normative behavior decreases with increasing costs.

The above utility function implies that the benefit of norm violation is homogeneous, i.e. norm targets receive the same payoff from norm violation in case their fraud is not detected. This is clearly a simplification of the reality: The benefit of doping heavily depends on an athlete’s physiological constitution whereas the benefit of tax fraud varies with the reported income. Nevertheless, this heterogeneity of the benefit of norm violation can be subsumed under the variable $c_i$ which originally represented norm target $i$’s initial adherence to the norm. Therefore, we refer to $c_i$ as the heterogeneous costs of norm violation. This extends the mechanism of Graetz et al. (1986) who assume that a certain fraction of taxpayers always reports their income truthfully independent of the economic incentives for tax evasion. Further, considering the aggregated benefit of (undetected) norm violation ($d_i = 1$) and the resulting disutility for the community, we obtain

$$r \sum_{d_i=1} b_a = r b_a n q = \sum_i r b_a q,$$

i.e. the aggregate benefit of norm violation multiplied by $r$ is identical to the aggregate disutility that norm violation imposes on others. For $r = 1$, norm violation induces a purely comparative advantage: If all norm targets violate the norm (i.e. $q = 1$), the benefit of fraud is zero, i.e. the comparative advantage vanishes if it refers to the whole population. This might apply to the example of doping whereas drug use (assuming a homogeneous effect on the athletes) is only effective if there are athletes who adhere to the anti-doping norm. With respect to tax evasion, we typically assume $r > 1$: The tax gap caused by an incorrect income report of a norm target cuts down the overall sum of taxes. Due
to economies of scale, the resulting decrease of investments in public goods outweigh the benefits of norm violation. In the absence of punishment, violating the norm always pays off at least for large populations as for each norm target, the disutility caused by her fraud is arbitrarily small. Thus, the only Nash equilibrium is the state where all norm targets violate the norm. As norm violation decreases welfare, this is a dilemma scenario and there is a demand to enforce the norm. One possible solution is altruistic punishment of norm violators by the norm targets. However, if detection is costly, inspection is often assumed by external institutions like the tax authority or the World Anti Doping Association.

There is no unique rule how these institutions choose the extent and quality of inspections. In the doping model of Berentsen (2002), each athlete is tested for drug use whereas the respective probability for each outcome of the inspection (correct positive/negative, false positive/negative) are exogenous and fixed. With respect to tax evasion, early models (Allingham and Sandmo, 1972; Yitzhaki, 1974) exhibit similar assumptions as there is an exogenous audit probability for taxpayers. Later extensions allow for an interaction between the taxpayers and the government by a variable taxation and auditing policy. Here, the tax authority typically maximizes expected net revenue (tax and penalty revenue, less audit costs). Some models distinguish between a government and its tax authority: While the latter maximizes expected net revenue, the government’s choice depends on a broad social welfare function. The interplay of these different incentive schemes and their effect on tax evasion is discussed in Melumad and Mookherjee (1989) and Cremer et al. (1990). Hence, the cost of an undetected tax gap is shifted from the individuals to the government in that case.

In our model, the inspector’s strategy only consists of the control effort \( e_c \). We assume that she gets a reward for each detected norm violator which constitutes an incentive for wide controls. Nevertheless, high expenditures on control effort will only pay off if the proportion of detected norm violators is sufficiently high so that the inspector’s utility also depends on the actual proportion of norm violators. Hence, for fixed concealment efforts \( e_a \), and a proportion of norm violators \( q \), the inspector’s expected utility is defined by

\[
u_c(e_c) = b_c \left( \frac{1}{n} \sum_{d_i=1}^{n} p(e_a, e_c, q) \right) - e_c.
\]

Hence, the inspector’s reward is proportional to the expected proportion of detected norm violators denoted by

\[\tilde{q} = \frac{1}{n} \sum_{d_i=1}^{n} p(e_a, e_c, q).\]

The reward for a detected norm violator is therefore \( \frac{b_c}{n} \). We can interpret \( u_c \) as the variable part of the inspector’s income whereas the fixed part does not influence the control effort.
4.1. A Public Goods Game with Unobservable Actions

In case that all norm targets violate the norm and are detected, the inspector achieves the maximum reward $b_c > 0$. As $u_c(0) \geq 0$, i.e. zero control effort always implies non-negative utility for the inspector, we obtain $u_c(e^*_c) \geq 0$ for the utility maximizing effort $e^*_c$. Hence, the inspector’s expenditures on control effort will not exceed the expected reward $\tilde{q} b_c$ if she maximizes her expected utility. This can be interpreted as an implicit budget constraint for the inspector.

The assumption of head money like incentives for effective controls is often valid with respect to criminal behavior.\(^4\) In accordance with Tsebelis (1990), detection of norm violators increases the inspector’s payoff while unsuccessful controls are costly. It is also similar to the model of tax evasion by Melumad and Mookherjee (1989) who discuss the transfer of authority over the income tax audit policy from the government to managers.

In the case of doping in sports this may not hold as there is a higher degree of interweavement of the norm targets and the inspection institution. The foundation of the World Anti Doping Association (WADA) by the International Olympic Committee is a first step to reduce this interweavement. The WADA is funded by public money and conducts a considerable proportion of doping tests (especially with respect to out-of-competition tests). In many sports, doping tests are conducted by the respective sports association which has no interest of convicting their athletes of doping. Additionally this could result in reduced income for all athletes and the association as sponsors might be deterred by a high rate of detected drug takers.

4.1.3 Rationality and Knowledge

With respect to the norm targets’ and the inspector’s knowledge, we make the following assumptions:

(K1) The utility functions $u_a$ of a norm target and $u_c$ of the inspector and the parameters $b_a$ (benefit of norm violation), $s$ (sanction cost) and $b_c$ (benefit of detection) are common knowledge.

(K2) The heterogeneous costs $c_i$ of norm violation are private information of norm target $i$. Further, neither any norm target nor the inspector know the distribution function $F$ of these costs.

(K3) The proportion $\tilde{q}(t)$ of detected norm violators at time $t$ is common knowledge at time $t + 1$.

\(^4\)Nevertheless, this implies incentives for deliberate false positives by the inspector. For historical examples with respect to taxes see Adams (1993).
(K4) The proportion $q(t)$ of norm violators at time $t$ cannot be observed by any norm target or the inspector at time $t + 1$ (or any other time step).

(K5) There are no false-positive detections of norm violators by the inspector.

We assume that norm targets and the inspector maximize their respective utility given the above knowledge constraints. All strategies, i.e. each norm target’s behavior and concealment effort on one hand and inspector’s control effort on the other, are chosen simultaneously at all time steps and cannot be observed afterwards. Due to our assumptions, the interaction between the norm targets and the inspector is partly strategic. For example, the norm targets know that the inspector’s benefit increases with the proportion of detected norm violators and account for this knowledge in their decision-making. On the other hand, the inspector and other norm targets are not able to foresee whether norm target $i$ violates the norm or not for given strategies of the other norm targets and the inspector as $c_i$ is unknown. Further, a prediction of norm violation is not only impossible on the individual level - also the overall extent of fraud captured by the proportion $q$ of norm violators in the population cannot be foreseen by the norm targets and the inspector as they do not know $F$. But as the inspector’s benefit and possibly also the detection probability depend on $q$, we assume that both parties have to estimate the proportion of norm violators $q(t)$ at time $t$ and use it within their decision-making process at the next time step. We denote these estimations by $\hat{q}_{a_i}(t)$ (norm targets) and $\hat{q}_c(t)$ (inspector).

If the distribution function $F$ of the heterogeneous costs was common knowledge, we could formulate our model as a Bayesian game whereas $F$ represents the norm targets’ and the inspectors prior beliefs about each norm target’s heterogeneous cost $c_i$. According to Harsanyi (1967-68), we could reformulate this game of incomplete information as a game of imperfect information and apply standard equilibrium concepts (see Fudenberg and Tirole, 1991, chapter 8). However, as we cannot identify specific prior beliefs in tax evasion or doping, we argue that our behavioral approach is more suitable to model the norm targets’ and the inspector’s decision-making. Further, our results do not depend on the actual shape of $F$.

The estimations of the real proportion of norm violators, which are then plugged in the utility functions to determine the resulting optimal decisions, are based on the proportion of detected norm violators $\bar{q}(t)$ at time $t$ that is common knowledge by (K3). This degree of information corresponds to the situation in the real world examples: Although nobody knows the extent of tax evasion or doping in the population, there is public information about how many taxpayers or athletes have been detected as norm violators in a certain period. With respect to doping, there is not only information about the extent of detected drug users among the athletes. Additionally, every detected norm violator’s identity is
published, whereas this is usually not the case for tax evasion. However, in our model, neither the norm targets nor the inspector take into account past norm violation on the individual level with respect to their decision-making. For example, the inspector does not apply higher inspection effort for norm targets that have violated the norm in previous periods. Note that as we excluded false positive detections by (K5), a high proportion of detected norm violators \( q \) induces more information about the real proportion of norm violators \( \dot{q} \) compared to a low proportion \( \ddot{q} \). The reason is that a low value of \( \ddot{q} \) might on the one hand result from a low value of \( q \). On the other hand, we may observe the same proportion of detected norm violators if there is a high proportion of norm violators paired with a low detection probability which can result from a high concealment effort or a low inspection effort. As \( \ddot{q} \) is only a lower bound for \( q \), all estimations \( q_a(t), q_c(t) \) have to fulfill

\[
\ddot{q}(t) \leq q_a(t), q_c(t) \leq 1.
\]

In general, an actual estimation by the norm targets or the inspector might depend on many factors. In Section 4.4.1 we specify a simple estimation procedure that combines objective information (\( \ddot{q} \)) with an individual’s (external) subjective belief about the extent of norm violation in the population.

Overall, we developed a model to analyze how the dark figure of norm violation influences the interaction between the norm targets and the inspector assuming that norm violation is only detected with a certain probability depending on the two parties’ respective effort. Both parties act partly strategically in the sense of knowing and considering the incentives of all involved individuals. However, they are only boundedly rational as certain parameters of their utility function are unobservable (the proportion of norm violators \( q \)) or private information (the heterogeneous cost \( c_i \)). In the following section, we specify further assumptions on the probability of detection ((P1)–(P7)) and the estimation procedure ((E1)–(E4)) and analyze the resulting dynamics. First, we investigate a scenario where the proportion of norm violators is observable at any time step. In Section 4.4.1 we analyze how the dynamics change if there is a dark field of norm violators, i.e. the proportion of norm violators is not observable.

### 4.2 Assumptions and Basic Properties

We model the interaction between the norm targets and the inspector as follows: At each time step, every norm target \( i \) chooses a strategy \( x_i(t) = (d_i(t), e_{ai}(t)) \) where \( e_{ai} \) denotes

\footnote{With respect to doping, Berry (2008) questions the validity of this assumption.}
the concealment effort and $d_i$ indicates whether norm violation pays off ($d_i = 1$) or not ($d_i = 0$). Simultaneously, the inspector chooses the control effort $e_c(t)$. Both parties maximize their respective expected utility according to Eqs. (4.3) and (4.4) on the basis of the observed consequences of their actions in the last period, namely the proportion of norm violators or detected norm violators. Within one time step, we assume strategic interaction between the norm targets and the inspector: Both parties have knowledge of each others utility function and can incorporate it in their respective optimization. Nevertheless, we restrict their strategic horizon to the actual time step by (K2): The norm targets and the inspector do not take possible actions in future periods into account as neither party has knowledge about the heterogeneous cost distribution $F$ (and therewith the actual proportion of norm violators) — a norm target $i$ only knows her own value $c_i$.

We now specify further assumptions with respect to the shape of the detection probability function $p$. First, we require that for every norm target, the marginal probability $\frac{\partial p}{\partial e_a}$ of being detected with respect to her effort $e_a$ does not depend on the inspector’s expenditures $e_c$ and vice versa. This leads to

\begin{equation}
(P1) \quad p(e_a, e_c, q) = (1 - \alpha) f_c(e_c, q) - \alpha f_a(e_a, q) + \alpha
\end{equation}

with $f_a, f_c : \mathbb{R}_+ \times [0, 1] \to [0, 1]$ denoting the effect of the respective effort on the probability of catching a norm violator. The parameter $\alpha = p(0, 0, q)$ measures the probability of being detected when both parties’ effort is zero. Hence, a low value of $\alpha$ leads to an initial advantage for the norm violators as the revelation probability is low in this situation. However, the marginal effect of every norm target’s effort is decreased by a low value of $\alpha$ so that her initial advantage disappears with increasing efforts. For the extreme example $\alpha = 0$ ($\alpha = 1$) detection is impossible (guaranteed) for zero efforts of both parties, while any effort of the norm targets (the inspector) has no effect. Considering the example of fare dodging, the probability of detecting a passenger without a ticket is mainly affected by the inspector’s effort which in this case represents the frequency of ticket examination. In contrast, the passenger’s ability to conceal his fraud in case of inspection is very limited which overall is reflected in $\alpha \approx 0$ in this situation.

Further, for $i = a, c$, $f_i$ has to fulfill the following requirements:

(P2) $f_i$ is continuous and twice continuously differentable on its interior.

(P3) $f_i(\cdot, q)$ is strictly concave for all $q$.

(P4) $f_i(0, q) = 0$ for all $q$.

(P5) $\lim_{e_i \to 0} \frac{\partial f_i}{\partial e_i}(e_i, q) = \infty$ for all $q$. 

While (P2) is a purely technical assumption, (P3) guarantees that the detection probability is strictly decreasing (strictly increasing) in the concealment (control) effort and that there is a decreasing marginal effect of both kinds of investments on the revelation probability of norm violation. (P4) reflects that without any investments, neither the inspector nor the norm targets can influence the probability $p$ of catching a norm violator. Finally, (P5) assures that the marginal effect of investments is infinitely high in case of zero expenditures. Hence, the inspector and a norm violator will always exhibit non-zero investments. The assumptions (P2)–(P5) basically guarantee that there is always a unique optimal concealment and control effort for any proportion of norm violators.

Additionally, we confine the dependence of the detection probability on the proportion of norm violators to effects caused by indirect detection:

(P6) \[ \frac{\partial f_a}{\partial q} \leq 0, \quad \frac{\partial f_c}{\partial q} \geq 0 \]

(P7) \[ \frac{\partial^2 f_a}{\partial e_a \partial q} \leq 0, \quad \frac{\partial^2 f_c}{\partial e_c \partial q} \geq 0 \]

By this assumptions, the effect and the marginal effect of expenditures on concealing a norm violation are non-increasing with the proportion of norm violators $q$ while the effect and the marginal effect of inspection investments do not decrease with $q$.

**Example 4.1.** An example for a function $p$ satisfying all the above assumptions is

\[
p(e_a, e_c, q) = (1 - \alpha) \left[ 1 - \exp \left( -e_c^{\beta} \right) \right] - \alpha \left[ 1 - \exp \left( -e_a^{\beta} \right) \right] (1 - \gamma q^2) + \alpha
\]

with parameters $\beta, \gamma \in (0, 1)$.

Here, the effect of the effort expenditures $e_a$ and $e_c$ on the probability of detecting a norm violator increases with $\beta$ while $\gamma$ determines the extent to which the effect of the concealment effort depends on the proportion of norm violators $q$. In Figure 4.2 we depict some properties of $p$ for $\alpha = \beta = \gamma = 0.5$. Note that we use this example only to illustrate certain properties of our model - all following results hold for any function $p$ fulfilling the above assumptions and do not depend on the particular choice of $p$.

Based on these assumptions, we now derive the dynamics for the key variable, namely the proportion of norm targets in the population who do not adhere to the norm.

---

6Strict concavity, domain $\mathbb{R}_+$ and a non-negative codomain imply that $f_i$ is strictly increasing.
Figure 4.2: The probability $p(e_a, e_c, q)$ of being detected dependent on the concealment effort $e_a$, the control effort $e_c$ and the proportion of norm violators $q$ for Example 4.1 with parameters $\alpha = \beta = \gamma = 0.5$. In (a) and (b), we depict $p$ as a function of $e_a$ (or $e_c$ respectively). With $q$ fixed, we choose different values of $e_c$ ($e_a$) whereas the arrow indicates the direction of increase. In (c), we fix $e_a$ and $e_c$ and depict $p$ as a function of $q$ where $p(e_a, e_c, \cdot)$ decreases with $e_a$ and increases with $e_c$. (d) shows $p$ as a function of $e_a$ and $e_c$ for $q$ fixed.
4.3 Equilibrium Properties for Observable Proportion of Norm Violators

First we analyze the model assuming that the proportion $q$ of norm violators is observable by the norm targets and the inspector. The inspector’s optimal effort $e^*_e(t)$ requires

$$ q(t) \frac{\partial f_c}{\partial e_c}(e^*_c(t + 1), q(t)) = \frac{1}{(1 - \alpha) b_c} $$

(4.7)

whereas (P3)–(P5) guarantee its existence and uniqueness for all parameters. In the following, $e^*_c(q(t))$ denotes the inspector’s optimal effort at time $t + 1$ for fixed $\alpha$ and $b_c$. Note that it also depends on $q(t)$ if the detection probability is independent of the proportion of norm violators.

Norm target $i$’s optimal effort $e^*_a(t + 1)$ has to fulfill

$$ \frac{\partial f_a}{\partial e_a}(e^*_a(t + 1), q(t)) = \frac{1}{\alpha s} $$

(4.8)

whereas for all parameters, the optimum always exists and is unique as (P3)–(P5) hold. As each norm target observes the same proportion of norm violators $q(t)$, the optimal concealment effort is identical for all norm targets. In the following, $e^*_a(q(t))$ denotes a norm target’s optimal effort at time $t + 1$. Note that this optimal effort is independent of $q(t)$ in case the detection probability is independent of the proportion of norm violators in the population. Additionally, the inspector’s decision influences whether the norm target actually violates the norm or not. Norm target $i$ will violate the norm at time $t + 1$ if

$$ u_{a_i}((0, 0)) < u_{a_i}\left((1, e^*_a(q(t)))\right) $$

or equivalently

$$ c_i < b_a - sp(e^*_a(q(t)), e^*_c(q(t)), q(t)) - e^*_a(q(t)). $$

(4.9)

Here, the norm targets are able to determine the inspector’s optimal effort as $q(t)$ and the inspector’s utility are common knowledge. Note further that a norm target’s decision whether to violate the norm or not does not depend on the factor $r$ and is thus not affected by the extent of welfare loss for the population caused by norm violation.

One can easily verify how the optimal efforts respond to a change in the proportion $q$ of norm violators (cf. Figure 4.3):

**Proposition 4.2.** For all $b_a, b_c, s \in \mathbb{R}_+, \alpha \in [0, 1]$,

(i) $e^*_a(q)$ is non-increasing in $q$

(ii) $e^*_c(q)$ is strictly increasing in $q$

for any proportion of norm violators $q \in [0, 1]$. 

Proof. From Eq. (4.8) and (4.7) we can conclude that

\[
\frac{\partial e^*_a}{\partial q} = -\frac{\partial^2 f_a}{\partial e_a \partial q} \leq 0,
\]

\[
\frac{\partial e^*_c}{\partial q} = -q \frac{\partial^2 f_c}{\partial e_c \partial q} + \frac{\partial f_c}{\partial e_c} > 0
\]

as \( f_a, f_c \) are strictly concave and we required (P6) and (P7).

Figure 4.3: Illustration of Proposition 4.2. An increase in the proportion of norm violators from \( q_1 \) to \( q_2 \) leads to a decrease of the optimal concealment effort \( e^*_a \) and an increase of the optimal control effort \( e^*_c \).

As both parties know \( q(t) \), they can immediately determine their respective optimal effort level. To check whether norm violation pays off, each norm target additionally can use her knowledge about the inspector’s utility to obtain \( e^*_c(t) \) and thus insert it in Eq. (4.9). Hence, the new proportion of norm violators

\[ q(t + 1) = F\left(b_a - s p(e^*_a(q(t)), e^*_c(q(t)), q(t)) - e^*_a(q(t))\right) =: g_q(q(t)). \]  

is completely determined by the respective proportion at the previous time step (see Figure 4.4). Any equilibrium proportion of norm violators \( q^* \) thus is a fixed point of \( g_q \), i.e. \( q^* = g_q(q^*) \). Further, we denote the proportion of detected norm violators in equilibrium by

\[ \bar{q}(q^*) = q^* \ p(e^*_a(q^*), e^*_c(q^*), q^*). \]
We know how the iteration function $g_q$ responds to a change in the proportion of norm violators:

**Proposition 4.3.** If $F$ is differentiable in all except finitely many points, the proportion of norm violators $q(t+1)$ at time $t+1$ decreases with the proportion of norm violators $q(t)$ at time $t$.

**Proof.** We have

$$g'_q(q) = F'(b_a - sp(e^*_a(q), e^*_c(q), q) - e^*_c(q)) c(q) \geq 0$$

with

$$c(q) = -s (1 - \alpha) \left( \frac{\partial f_c}{\partial e_c} (e^*_c(q), q) \frac{\partial e^*_c}{\partial q} (q) + \frac{\partial f_c}{\partial q} (e^*_c(q), q) \right)$$

$$+ s \alpha \left( \frac{\partial f_a}{\partial e_a} (e^*_a(q), q) \frac{\partial e^*_a}{\partial q} (q) + \frac{\partial f_a}{\partial q} (e^*_a(q), q) \right) - \frac{\partial e^*_a}{\partial q} (q)$$

$$= -s \left[ \frac{1}{q b_c} \frac{\partial e^*_c}{\partial q} (q) + (1 - \alpha) \frac{\partial f_c}{\partial q} (e^*_c(q), q) - \alpha \frac{\partial f_a}{\partial q} (e^*_a(q), q) \right] \leq 0$$

and thus $g'_q(q) \leq 0$ if $g_q$ is differentiable at $q$ by using Proposition 4.2, assumptions (P6) and (P7), Eqs. (4.8) and (4.7). Hence, $g_q$ is decreasing as it is continuous in the finitely many points in which it is not differentiable.

Note that by allowing a finite number of points where $F$ is not differentiable we do not exclude for instance a uniform distribution on an interval. Further, the proof shows that our assumptions (P6) and (P7) are necessary for the above proposition as the effect of an increase of $q(t)$ on the detection probability might be ambiguous if one of these assumptions did not hold.

We now show that this equilibrium is unique for arbitrary parameters and derive its comparative statics.

**Theorem 4.4.** If $F$ is differentiable in all except finitely many points, $g_q$ has a unique fixed point $q^* (b_a, s, b_c, F)$ which is

(i) increasing in $b_a$,

(ii) decreasing in $s$ and $b_c$

for all parameters $b_a, s, b_c > 0$. Further, the associated proportion of detected norm violators $\tilde{q}(q^*(b_a, s, b_c, F))$ in equilibrium is increasing in $b_a$. 


Figure 4.4: Dynamics for an observable proportion of norm violators. A plus (minus) sign indicates that an increase of the respective input variable causes an increase (decrease) in the dependent variable whereas the monotonicity is not strict. The dotted line represents the multi-level relation between the proportion of norm violators at time $t$ and the proportion of norm violators at time $t+1$.

**Proof.** We have to show that $g_q$ has a unique fixed point. The Brouwer fixed point theorem immediately guarantees the existence of a fixed point $q^*$. The uniqueness of $q^*$ is ensured by the fact that $g_q$ is decreasing according to Proposition 4.3.

As $g_q$ is decreasing for all $b_a, s, b_c$, monotonicity of $g_q$ in one of this parameters implies a corresponding monotonicity of $q^*(b_a, s, b_l)$ in that parameter. One can verify that $\frac{\partial g_q}{\partial b_a} \geq 0$, $\frac{\partial g_q}{\partial s}, \frac{\partial g_q}{\partial b_c} \leq 0$ (where $F$ is differentiable) which proves the monotonicity properties of $q^*$. For $\tilde{q}^* = q^* p(e^*_a(q^*), e^*_c(q^*), q^*)$ we obtain

$$\frac{\partial \tilde{q}^*}{\partial b_a} = \frac{\partial q^*}{\partial b_a} p(e^*_a(q^*), e^*_c(q^*), q^*) + q^* \left( \frac{\partial p}{\partial e_a} \frac{\partial e^*_a(q^*)}{\partial q^*} + \frac{\partial p}{\partial e_c} \frac{\partial e^*_c(q^*)}{\partial q^*} + 1 \right) \geq 0$$

where $F$ is differentiable.

Hence, the assumption of an observable proportion of norm violators leads to a unique equilibrium $q^*$ which increases with the benefit of norm violation $b_a$ and decreases with the sanction cost $s$ for a detected norm violator and the inspector’s maximum reward $b_c$ for detected norm violation. Note that this holds for all parameters and in particular for any distribution of heterogeneous costs. We can also predict that the proportion of detected norm violators $\tilde{q}^*$ in equilibrium increases with $b_a$ as this implies more norm violators and a higher detection probability. In contrast, the effect of an increase of $s$ or $b_c$ on $\tilde{q}^*$ is ambiguous: The proportion of norm violators is reduced in both cases, but the increased quality of inspections counteracts this effect and prohibits any general prediction.

For a finite number of norm targets, the assumption of a continuous heterogeneous cost distribution is obviously violated. Here, the equilibrium may not be reached as the fraction of drug users is always rational in this situation. Nevertheless, one can interpret the equilibrium $q^*$ as the expected proportion of norm violators when the heterogeneous cost
4.3. Equilibrium Properties for Observable Proportion of Norm Violators

Figure 4.5: Dynamics of the proportion of norm violators $q$, the proportion of detected norm violators $\tilde{q}$ and the concealment/control effort $e_a/e_c$ for Example 4.1 (including cobweb diagram for $q$) with $b_u = b_c = 20$, $c_i$ uniformly distributed on $[0, b_a]$ and $\alpha = 0.2$, $\beta = 0.8$, $\gamma = 0.5$ whereas initially all norm targets are norm adherent. In (a), the equilibrium is stable, while an increase in the punishment $s$ for detected norm violators leads to a destabilization of the fixed point in (b). Note that the optimal concealment effort can be positive although all norm targets adhere to the norm.

The equilibrium is not necessarily stable, i.e. for an initial proportion of norm violators which is different from $q^*$ convergence is not guaranteed. For Example 4.1, dependent on the parameters one can also observe alternating proportions of norm violators. In Figure 4.5, we compare two scenarios whereas the parameters only differ in the punishment $s$ for detected norm violators. For a low value of $s$, we observe convergence towards the equilibrium proportion of norm violators $q^*$. An increase of $s$ leads to destabilization...
as $q$ alternates between zero and a positive value. Here, there is initially no incentive for controls as all norm targets adhere to the norm. The norm targets anticipate this behavior, and norm violation pays off for a non-zero proportion of the population. At the next time step, this can be observed by the inspector who therefore increases the control effort. Due to the high punishment of norm violation, this is sufficient to make all norm targets refrain from norm violation and we will hence observe oscillating behavior instead of convergence towards the equilibrium.

As we did in general not specify the function of probability of detection $p$ in detail, we cannot determine specific ranges within the parameter space that yield a stable equilibrium for arbitrary $p$. Nevertheless, Eq. (4.11) shows that $|g'(q)|$ increases with $s$ and decreases with $b_c$. Thus, increasing punishment for detected norm violators destabilizes the system whereas an increase of the inspector’s benefits has the opposite effect.\footnote{For any fixed point $q^*$ of $g_q$, $|g'_q(q^*)| < 1$ implies stability.}

4.4 Unobservable Proportion of Norm Violators

So far we assumed that at any time step, the norm targets and the inspector know the actual proportion of norm violators $q$ and therefore are able to determine their optimal decisions based on that value. In many situations this information is not available for either party, e.g. in the context of tax fraud or doping in sports. In order to determine their optimal effort at time $t+1$ according to Eqs. (4.8) and (4.7) respectively, they need to estimate $q(t)$ based on the public information of the proportion of detected norm violators $\tilde{q}(t)$ in the previous period. Any estimation has to be consistent with that information: If the proportion of detected norm violators at time $t$ is $\tilde{q}(t)$, a norm targets’ estimation $\hat{q}_a(t)$ and the inspector’s estimation $\hat{q}_c(t)$ of the real proportion of norm violators $q(t)$ at that time naturally have to fulfill Eq. (4.6) as this always holds for the estimated value $q$ in absence of false-positive detections.

4.4.1 Estimation of the Dark Field of Norm Violators

The norm targets and the inspector now have to choose a feasible estimation of $q$ according to Eq. (4.6). We assume that this choice is determined by the moral confidence, i.e. the belief of the extent of norm compliance in the population. In our model, a norm target’s moral confidence is measured by the suspiciousness $\lambda_a \in [0,1]$ which denotes the antonym (i.e. the belief of the extent of norm violation). The inspector’s suspiciousness towards the extent of norm violation is measured by the parameter $\lambda_c \in [0,1]$. Note that these
parameters are exogenous and not explained by our model. Additional requirements to
the estimation procedure are the following:

(E1) A norm target’s (or the inspector’s) estimation $\hat{q}_a(t)$ ($\hat{q}_c(t)$) of the real proportion
of norm violators is increasing in the proportion of detected norm violators $\tilde{q}(t)$ for
any given suspiciousness $\lambda_a$ ($\lambda_c$) at any time step $t$.

(E2) A norm target’s (or the inspector’s) estimation $\hat{q}_a(t)$ ($\hat{q}_c(t)$) of the real proportion
of norm violators is increasing in the suspiciousness $\lambda_a$ ($\lambda_c$) for any given proportion
of detected norm violators $\tilde{q}(t)$ at any time step $t$.

(E3) The suspiciousness of a norm target (of the inspector) is private information and
does not change over time.

For the sake of simplicity, we assume that the estimation of the real proportion of norm
violators is an affine-linear function of the suspiciousness and the proportion of detected
norm violators respectively:

$$\hat{q}_j(t) = \tilde{q}(t) + \lambda_j (1 - \tilde{q}(t)), \quad j = a, c.$$ (4.12)

For $\lambda_a = 0$ ($\lambda_a = 1$), norm target $i$ assumes the minimum (maximum) control level, i.e.
$\hat{q}_a = 1$ ($\hat{q}_a = \tilde{q}$). For $\lambda_a = 0.5$, the norm target’s estimation is the arithmetic mean of
the extreme values (analogously for $\lambda_c$). $\lambda_a$ ($\lambda_c$) is also the estimated proportion of norm
violators of a norm target (the inspector) if no norm violators are detected. Hence, the
norm targets and the inspector combine private beliefs (which are exogenous and constant)
and public information (which is explained by the model) to estimate the proportion of
norm violators. By (E3), we bound the norm targets’ and the inspector’s rationality as
any individual (norm target or inspector) is only aware of her own suspiciousness and
therefore cannot foresee other individuals’ estimations of the proportion of norm violators
$q$. Instead we assume that she plugs in her own estimation in case the decision-making
requires another individual’s estimation of $q$ (e.g. a norm targets has to consider the
inspector’s incentives which depends on the proportion of norm violators). Additionally,
we require that the suspiciousness is homogeneous among the norm targets:

(E4) The suspiciousness of all norm targets is identical, i.e. $\lambda_{a_i} = \lambda_a$.

Hence, all estimations of the real proportion of norm violators by the norm targets are
identical, i.e. $\hat{q}_{a_i} = \hat{q}_a$.

---

8Without changing the results one could also assume any function $z : [0, 1] \times [0, 1] \rightarrow [\tilde{q}, 1]$ that is
increasing in both arguments to locate the estimation $\hat{q} = z(\tilde{q}, \lambda)$ of $q$ in a more general way.
Figure 4.6: Dynamics of the proportion of detected norm violators $\tilde{q}$, the proportion of norm violators $q$ and the estimated proportion of norm violators $\hat{q}_a$ and $\hat{q}_c$ if the proportion of norm violators is unobservable for Example 4.1 (including cobweb diagram for $\tilde{q}$) with $\lambda_a = 0.05$, $\lambda_c = 0.3$, $b_a = b_c = s = 20$, $c_i$ uniformly distributed on $[0, b_a]$ and $\alpha = 0.2$, $\beta = 0.8$, $\gamma = 0.5$ whereas initially all norm targets are norm adherent.

Note that in principle, the norm targets and the inspector can shorten the interval of feasible estimations by incorporating the knowledge of their optimization procedure and their actions at past time steps. See Section 4.5 for a discussion.

4.4.2 Equilibrium Properties

With these assumptions, the dynamics of our model in case of an unobservable proportion of norm violators can now be rewritten in terms of the proportion of detected norm violators $\tilde{q}$. With $\tilde{q}(t)$ denoting this proportion at time $t$, we obtain

$$\tilde{q}(t + 1) = q(t + 1) p\left(e_a^*(\hat{q}_a(t)), e_c^*(\hat{q}_c(t)), q(t + 1)\right) =: g_\tilde{q}(\tilde{q}(t); \lambda_a, \lambda_c) \tag{4.13}$$

where $\hat{q}_a(t) = \hat{q}_a(\tilde{q}(t))$ and $\hat{q}_c(t) = \hat{q}_c(\tilde{q}(t))$ depend on $\tilde{q}(t)$, $\lambda_a$, $\lambda_c$ via Eqs. (4.6) and (4.12) and $q(t + 1) = g_q(\hat{q}_a(t))$. Note that a fixed point $\tilde{q}^*$ of $g_\tilde{q}$ induces an equilibrium proportion of norm violators that we denote by $q^*(\tilde{q}^*)$. Figure 4.6 depicts the dynamics according to Eq. (4.13) for Example 4.1.

Further, we can specify how the iteration function $g_\tilde{q}$ depends on the parties' respective
4.4. Unobservable Proportion of Norm Violators

suspiciousness:

**Proposition 4.5.** For all $\tilde{q}, \lambda_a, \lambda_c \in [0, 1],$

1. $g_{\tilde{q}}(\tilde{q}, \lambda_a, \lambda_c)$ is increasing in $\lambda_c,$
2. $q(\tilde{q})$ is decreasing in $\tilde{q}$

where $q(\tilde{q})$ denotes the proportion of norm violators resulting from a proportion $\tilde{q}$ of detected norm violators.

**Proof.** First, the norm targets’ estimation $\hat{q}_a$ of the proportion of norm violators does not depend on $\lambda_c$. Hence, we obtain

$$\frac{\partial g_{\tilde{q}}}{\partial \lambda_c}(\tilde{q}, \lambda_a, \lambda_c) = q(\hat{q}_a) \frac{\partial p}{\partial e_c}(e_a(\hat{g}_a), e_c(\hat{q}_c), q(\hat{q}_a)) \frac{\partial e_c^*}{\partial \hat{q}_c}(\hat{q}_c) \frac{\partial \hat{q}_c}{\partial \lambda_c}(\tilde{q}, \lambda_c) \geq 0$$

as $p$ increases with $e_c^*$, $e_c^*$ increases with $\hat{q}_c$ (cf. Proposition 4.2(ii)) and

$$\frac{\partial \hat{q}_c}{\partial \lambda_c}(\tilde{q}, \lambda_c) = (1 - \tilde{q}) \geq 0.$$

Second, $q(\tilde{q}) = g_{\tilde{q}}(\hat{q}_a(\tilde{q}))$ leads to

$$\frac{dg_{\tilde{q}}}{d\tilde{q}}(\hat{q}_a(\tilde{q})) = \frac{dg_{\tilde{q}}}{d\tilde{q}}(\hat{q}_a)(1 - \tilde{q}) \leq 0$$

as $\frac{dg_{\tilde{q}}}{d\tilde{q}} \leq 0$ according to Proposition 4.3. □

After having introduced uncertainty by the unobservability of the proportion of norm violators, the first question is whether Theorem 4.4 still holds for the new dynamics, i.e. if there is always a unique equilibrium proportion of norm violators. The answer is no: In Figure 4.7 we show that there are parameter settings where multiple fixed points of $g_{\tilde{q}}$ can arise in Example 4.1 given a dark figure of norm violation. In this example, the number of fixed points depends on the inspection reward $b_c$: for $0 < b_c < b_c^1 \approx 6.4$, there is a unique stable fixed point of $g_{\tilde{q}}$ and therefore also a unique stable equilibrium proportion of norm violators. For $b_c^1 < b_c < b_c^2 \approx 12.2$, we observe two stable equilibria: one where approximately 96% of the norm targets violate the norm and one where the proportion of norm violators varies, depending on $b_c$, between 40% and 70%. If $b_c^2 < b_c$, there is again a unique equilibrium.

The reason for this is that $g_{\tilde{q}}$ is not necessarily decreasing in $\tilde{q}$ (see Figure 4.8): An increase of $\tilde{q}$ will also increase the parties’ estimation $\hat{q}_a$ and $\hat{q}_c$ of the real proportion of

\[9\] The Brouwer fixed point theorem guarantees the existence of an equilibrium.
Figure 4.7: Example 4.1 with $b_a = s = 10$, $\lambda_a = \lambda_c = 0.01$, $c_i$ uniformly distributed on $[0, b_a]$ and $\alpha = 0.04$, $\beta = 0.9$, $\gamma = 0.1$. (a) depicts how the equilibrium proportion of norm violators change with the detection reward $b_c$ for observable (left) and unobservable (right) proportions of norm violators $q(t)$. A plus indicates stable, a circle unstable equilibria. In (b), we show the equilibrium proportions of detected norm violators (the fixed points of $g_{\bar{q}}$) for unobservable $q(t)$ and the iteration function $g_{\bar{q}}$ for $b_c = 2$ (one fixed point), $b_c = 8$ (three fixed points: two stable, one unstable) and $b_c = 20$ (one fixed point).
norm violators $q$ (cf. Eq. (4.6), (4.12)). According to Proposition 4.2, this leads to an increase of $e^*_c$, a decrease of $e^*_a$ and in addition to a decrease of the resulting proportion of norm violators $q$ as $g_q$ is decreasing (with argument $\hat{q}_a$). However, the probability of being detected decreases in this situation. Hence, we obtain fewer norm violators but also a lower probability of being detected. Consequently we cannot make general predictions about the new proportion of detected violators.

Figure 4.8: Dynamics for an unobservable proportion of norm violators. A plus (minus) sign indicates that an increase of the respective input variable causes an increase (decrease) in the dependent variable whereas the monotonicity is not strict. A question mark denotes that the effect of a change in the input variable cannot be predicted in general. Multi-level relations as the response of the proportion of norm violators to a change in the proportion of detected norm violators are represented by a dotted line. The arrow label indicates the origin of the respective relation.

### 4.4.3 An Inverse Self-Fulfilling Prophecy Effect

Our basic research question is how estimation of the dark figure of norm deviance influences the actual strength of a norm. Regarding our model, we can reformulate this question: How does a change in the norm targets’ and the inspector’s estimation procedure of the real proportion of norm violators, i.e. a change in $\lambda_a$ or $\lambda_c$, influence the proportion of norm violators in equilibrium? As we only made very basic assumptions about the dynamics by not specifying the dependence of the probability of detecting norm violation on the parties’ respective effort and the proportion of norm violators in detail, an all-encompassing answer is not possible. Nevertheless we are able to make very general predictions in Theorem...
4.8 about how a “typical” stable equilibrium $\tilde{q}^\ast$ (i.e. $\tilde{q}^\ast$ is asymptotically stable and hyperbolic\textsuperscript{10}) of detected norm violators will respond to small changes in the suspiciousness parameters $\lambda_a$ and $\lambda_c$ of the norm targets and the inspector. There, we make use of the following lemmas referring to that type of fixed point.

**Lemma 4.6.** Let $x_{t+1} = f(x_t)$ denote a difference equation with $f : [0, 1] \to [0, 1]$ continuous. If $x^\ast$ is an asymptotically stable fixed point of $f$, there is $\delta > 0$ with

1. $f(x) < x$ if $x \in (x^\ast, x^\ast + \delta)$
2. $f(x) > x$ if $x \in (x^\ast - \delta, x^\ast)$

Proof. If for all $\delta > 0$ there is $x \in (x^\ast, x^\ast + \delta)$ with $f(x) = x$, $x^\ast$ cannot be attractive and thus not asymptotically stable. Hence let us assume that we always find $x$ with $f(x) > x$ in the same interval. As $f$ is continuous, this implies that there is $\varepsilon > 0$ with

$$f(x) > x \text{ for } x \in [x^\ast, x^\ast + \varepsilon].$$

Then $x^\ast$ cannot be stable: for all $\delta' > 0$ (with $\delta' < \varepsilon$) we can choose $x_0 \in (x^\ast, x^\ast + \delta')$. If we assume $|x_t - x^\ast| = x_t - x^\ast < \varepsilon$ for all $t$, $(x_t)_t$ must converge with

$$x^\ast < \lim_{t \to \infty} x_t \leq \varepsilon.$$

But then the limit is a fixed point of $f$:

$$f(\lim_{t \to \infty} x_t) = \lim_{t \to \infty} f(x_t) = \lim_{t \to \infty} x_{t+1} = \lim_{t \to \infty} x_t$$

as $f$ is continuous. This is a contradiction to Eq. (4.14), hence we proved (i). The proof for (ii) is analogous. \qed

**Lemma 4.7.** Let $f : [0, 1] \times [0, 1] \to [0, 1]$ be continuous and let $x_{t+1} = f_\lambda(x_t)$ denote a difference equation with $f_\lambda := f(\cdot, \lambda)$ continuously differentiable for all $\lambda$. If $x^\ast(\lambda)$ is an asymptotically stable hyperbolic fixed point of $f_\lambda$, then for all $\delta_x > 0$ there is $\delta_\lambda > 0$ so that for all $\lambda' \in [0, 1]$ with $|\lambda' - \lambda| < \delta_\lambda$ there is a unique fixed point $\tilde{x}$ of $f_{\lambda'}$ in $[x^\ast - \delta_x, x^\ast + \delta_x]$, i.e.

$$|\lambda' - \lambda| < \delta_\lambda \Rightarrow \exists \tilde{x} \in [x^\ast - \delta_x, x^\ast + \delta_x] \text{ with } f_{\lambda'}(\tilde{x}) = \tilde{x} \text{ and } |f'_{\lambda'}(\tilde{x})| < 1.$$

Moreover, $\tilde{x}$ is asymptotically stable and hyperbolic.

\textsuperscript{10}See e.g. Elaydi (1996) for a definition of an asymptotically stable and hyperbolic equilibrium.
4.4. Unobservable Proportion of Norm Violators

Proof. According to Lemma 4.6 and the fact that \( x^* \) is an asymptotically stable hyperbolic fixed point, we can find \( \delta > 0 \) (\( \delta \leq \delta_x \)) with

\[
f_{\lambda}(x) = \begin{cases} 
  x & \text{if } x \in [x^*, x^* + \delta] \\
  > x & \text{if } x \in [x^* - \delta, x^*] 
\end{cases}
\tag{4.15}
\]

and

\[|f'_{\lambda}(x)| < 1 \quad \text{for } x \in I\]

whereas \( I := [x^* - \delta, x^* + \delta] \). As \( f_{\lambda} \) is continuous on the compact set \( I \), we can define

\[\varepsilon_1 := 1 - \max_{x \in I} |f'_{\lambda}(x)| > 0.\]

Further we can choose \( x_-, x_+ \in I \) with \( x_- < x^* \) and \( x_+ > x^* \) and define

\[\varepsilon_2 := \min(f_{\lambda}(x_-) - x_-, x_+ - f_{\lambda}(x_+)) > 0\]

according to Eq. (4.15). As \( \frac{\partial f}{\partial x} \) is uniformly continuous on the compact set \( I \times [0, 1] \), there is \( \delta_1 > 0 \) so that for \( (x_1, \lambda_1), (x_2, \lambda_2) \in I \times [0, 1] \)

\[\| (x_1, \lambda_1) - (x_2, \lambda_2) \| < \delta_1 \Rightarrow |f'_{\lambda_1}(x_1) - f'_{\lambda_2}(x_2)| < \varepsilon_1\]

holds whereas \( \| \cdot \| \) denotes the Euclidean norm. In particular, we thereby obtain

\[|\lambda' - \lambda| < \delta_1 \Rightarrow |f'_{\lambda'}(x) - f'_{\lambda}(x)| < \varepsilon_1\]

for all \( x \in I \) and \( \lambda' \in [0, 1] \) as \( \| (x, \lambda') - (x, \lambda) \| = |\lambda' - \lambda| \).

Further, the continuity of \( f(x, \cdot) \) for all \( x \) guarantees the existence of \( \delta_2 > 0 \) with

\[|\lambda' - \lambda| < \delta_2 \Rightarrow |f_{\lambda'}(x_-) - f_{\lambda}(x_-)|, |f_{\lambda'}(x_+) - f_{\lambda}(x_+)| < \varepsilon_2\]

for \( \lambda' \in [0, 1] \). Overall, this implies

\[|\lambda' - \lambda| < \delta_\lambda \Rightarrow |f'_{\lambda'}(x)| < 1 \quad \text{for all } x \in I, \quad f_{\lambda'}(x_-) - x_- > 0, \quad f_{\lambda'}(x_+) - x_+ < 0\]

with \( \delta_\lambda := \min(\delta_1, \delta_2) \). Hence, according to the intermediate value theorem, we obtain

\[|\lambda' - \lambda| < \delta_\lambda \Rightarrow \exists \tilde{x} \in (x_-, x_+) \text{ with } f_{\lambda'}(\tilde{x}) = \tilde{x} \text{ and } |f'_{\lambda'}(\tilde{x})| < 1,\]

i.e. for small variations of \( \lambda \) there is always an asymptotically stable hyperbolic fixed point near the old equilibrium \( x^* \). Moreover, as \( |f'_{\lambda'}(x)| < 1 \quad |\lambda' - \lambda| < \delta_\lambda \text{ and } x \in I \), we know that \( \frac{\partial}{\partial x} f_{\lambda'}(x) - x < 0 \), i.e. \( \tilde{x} \) is the only fixed point of \( f_{\lambda'} \) on \( I \) for \( |\lambda' - \lambda| < \delta_\lambda \). \( \square \)
If we observe convergence towards an equilibrium \( \bar{q}^* \), then this fixed point is typically asymptotically stable and hyperbolic. Nevertheless we cannot exclude possible exceptions such as semi-stable or asymptotically stable but non-hyperbolic equilibria (or starting in an unstable equilibrium). But for any given functional form of \( p(e_a, e_c, q) \) it can easily be verified whether an equilibrium \( \bar{q} \) of \( g_q \) has the designated property.\(^{11}\)

**Theorem 4.8.** Let \( \bar{q}^*(\lambda_a, \lambda_c) \) denote an asymptotically stable hyperbolic fixed point of \( g_q \) for suspiciousness parameters \( \lambda_a, \lambda_c \in [0, 1] \) and let \( g_q \) be continuously differentiable at \( \bar{q}^*(\lambda_a, \lambda_c) \). Then there is \( \delta > 0 \) with

\[
\bar{q}^*(\lambda_a, \lambda'_c) \begin{cases} 
\leq \bar{q}^*(\lambda_a, \lambda_c) & \text{if } \lambda'_c \in [\lambda_c - \delta, \lambda_c] \\
\geq \bar{q}^*(\lambda_a, \lambda_c) & \text{if } \lambda'_c \in [\lambda_c, \lambda_c + \delta],
\end{cases}
\]

whereas \( \bar{q}^*(\lambda_a, \lambda'_c) \) denotes the unique fixed point of \( g_q \) on \([x^* - \delta, x^* + \delta]\) for suspiciousness parameters \( \lambda_a, \lambda'_c \).

**Proof.** We define \( h_{\lambda'_c}(\bar{q}) := g_q(\bar{q}, \lambda_a, \lambda'_c) \) for \( \lambda'_c \in [0, 1] \). Lemma 4.6 guarantees the existence of \( \delta_{\bar{q}} > 0 \) with

\[
h_{\lambda'_c}(\bar{q}) - \bar{q} \begin{cases} 
> 0 & \text{if } \bar{q} \in \left[\bar{q}^*(\lambda_a, \lambda_c) - \delta_{\bar{q}}, \bar{q}^*(\lambda_a, \lambda_c)\right] \\
< 0 & \text{if } \bar{q} \in \left[\bar{q}^*(\lambda_a, \lambda_c), \bar{q}^*(\lambda_a, \lambda_c) + \delta_{\bar{q}}\right]
\end{cases}
\] (4.16)

According to Lemma 4.7, there is further \( \delta_{\lambda'_c} > 0 \) so that for \( |\lambda'_c - \lambda_c| < \delta_{\lambda'_c} \) there is a unique fixed point of \( h_{\lambda'_c} \) in \([\bar{q}^*(\lambda_a, \lambda_c) - \delta_{\bar{q}}, \bar{q}^*(\lambda_a, \lambda_c) + \delta_{\bar{q}}]\) which we refer to as \( \bar{q}^*(\lambda_a, \lambda'_c) \). If \( \lambda'_c \in [\lambda_c - \delta_{\lambda'_c}, \lambda_c] \), Proposition 4.5 implies that \( h_{\lambda'_c}(\bar{q}) \leq h_{\lambda'_c}(\bar{q}) \) for \( \bar{q} \in [0, 1] \), which leads to

\[ h_{\lambda'_c}(\bar{q}) - \bar{q} \leq h_{\lambda'_c}(\bar{q}) - \bar{q} < 0 \text{ for } \bar{q} \in (\bar{q}^*(\lambda_a, \lambda_c), \bar{q}^*(\lambda_a, \lambda_c) + \delta_{\bar{q}}].\]

Hence, there is no fixed point of \( h_{\lambda'_c} \) in that interval and we thereby obtain

\[ \bar{q}^*(\lambda_a, \lambda'_c) \in \left[\bar{q}^*(\lambda_a, \lambda_c) - \delta_{\bar{q}}, \bar{q}^*(\lambda_a, \lambda_c)\right].\]

Analogously one can show that

\[ \bar{q}^*(\lambda_a, \lambda'_c) \in \left[\bar{q}^*(\lambda_a, \lambda_c), \bar{q}^*(\lambda_a, \lambda_c) + \delta_{\bar{q}}\right] \]

for \( \lambda'_c \in [\lambda_c, \lambda_c + \delta_{\lambda'_c}] \). \qed

\(^{11}\)|\(g_q(\bar{q}^*)| < 1\) is a sufficient condition.
This means that a small increase in the inspector’s suspiciousness $\lambda_c$ leads to an increase of the equilibrium proportion of detected norm violators $\tilde{q}^*$. The reason is that the iteration function $g_{\tilde{q}}$ increases with $\lambda_c$ as for a given proportion of detected norm violators, the inspector’s estimation of the proportion of norm violators increases with $\lambda_c$ (cf. Figure 4.9(a)). By this shift of $g_{\tilde{q}}$, the equilibrium proportion of detected norm violators must increase. Note that this conclusion is possible if the old equilibrium is asymptotically stable and hyperbolic and if the change in $\lambda_c$ is sufficiently small. As the equilibrium proportion of norm violators $q^*$ decreases with the norm targets’ estimation $\hat{q}_a^*$ which is positively affected by the increase of $\tilde{q}^*$ (cf. Figure 4.8), this will finally lead to a decrease of the proportion of norm violators $q^*$ in the new equilibrium.

**Corollary 4.9.** Let $\tilde{q}^*(\lambda_a, \lambda_c)$ denote an asymptotically stable hyperbolic fixed point of $g_{\tilde{q}}$ for suspiciousness parameters $\lambda_a, \lambda_c \in [0, 1]$ and let $g_{\tilde{q}}$ be continuously differentiable at $\tilde{q}^*(\lambda_a, \lambda_c)$. Let further $q^*(\lambda_a, \lambda_c) = q(\tilde{q}^*(\lambda_a, \lambda_c))$ denote the proportion of norm violators in that equilibrium. Then there is $\delta > 0$ with

\[
q^*(\lambda_a, \lambda_c) \begin{cases} 
\geq q^*(\lambda_a, \lambda_c) & \text{if } \lambda_c' \in [\lambda_c - \delta, \lambda_c] \\
\leq q^*(\lambda_a, \lambda_c) & \text{if } \lambda_c' \in [\lambda_c, \lambda_c + \delta],
\end{cases}
\]

whereas $\tilde{q}^*(\lambda_a, \lambda_c')$ denotes the unique fixed point of $g_{\tilde{q}}$ on $[x^* - \delta, x^* + \delta]$ for suspiciousness parameters $\lambda_a, \lambda_c'$.

**Proof.** The statement is an immediate consequence of Theorem 4.8 and the fact that

\[
\frac{\partial q^*}{\partial \lambda_c}(\lambda_a, \lambda_c) = \frac{\partial}{\partial \lambda_c} g(\hat{q}_a^*(\lambda_a, \lambda_c)) = g'(\hat{q}_a^*(\lambda_a, \lambda_c)) (1 - \lambda_a) \frac{\partial \hat{q}_a^*}{\partial \lambda_c}(\lambda_a, \lambda_c) \leq 0
\]

by using Proposition 4.3. \qed

Note that in general, we can only guarantee the existence of the threshold $\delta$ for a change in the inspector’s suspiciousness where a prediction regarding the change in the equilibrium proportion of norm violators is possible. Depending on the functional form of $p(e_a, e_c, q)$, the threshold could be arbitrarily small, whereas in other scenarios this prediction is possible for any change in the inspectors suspiciousness. Further, Theorem 4.8 might not be valid in case of a more sophisticated estimation procedure of the proportion of norm violators by the norm targets and the inspector. For example, both types of individuals could use their knowledge of their past actions or their counterpart’s estimation procedure and thereby shorten the interval of feasible estimations of the proportion of norm violators. One can show that this more sophisticated procedure does not fulfill assumption (E1) for arbitrary detection probability functions $p$. 

Figure 4.9: Change in the norm targets’ and the inspector’s suspiciousness for Example 4.1 with $b_a = b_c = s = 10$, $c_i$ uniformly distributed on $[0, b_a]$ and $\alpha = \beta = 0.9$, $\gamma = 0.5$. The left column depicts the iteration function $g_{\tilde{q}}$, the right column the equilibrium proportion of norm violators $q^*$ and detected norm violators $\tilde{q}^*$ depending on the respective suspiciousness. In (a) we increase the inspector’s suspiciousness $\lambda_c$ from 0.1 to 0.9 while in (b) the norm targets’ suspiciousness $\lambda_a$ is increased.

Without further requirements on $p$ it is not possible to deduce a similar result with respect to a change in the norm targets’ suspiciousness $\lambda_a$: It is not clear whether an increase of $\lambda_a$ increases or decreases the equilibrium proportion of detected norm violators $\tilde{q}^*$ (cf. Figure 4.9). The reason is that the iteration function $g_{\tilde{q}}$ is not necessarily decreasing in
4.5. Robustness of the Results

However, regarding the dependence of the equilibrium proportion of norm violators \( q^* \) on \( \lambda_a \), we obtain

\[
\frac{\partial q^*}{\partial \lambda_a}(\lambda_a, \lambda_c) = \frac{\partial}{\partial \lambda_a} g(q^*_a(\lambda_a, \lambda_c)) = g'(q^*_a(\lambda_a, \lambda_c)) \left[ (1 - \lambda_a) \frac{\partial q^*}{\partial \lambda_a}(\lambda_a, \lambda_c) + 1 - q^*(\lambda_a, \lambda_c) \right].
\]

Here, \( 1 - q^*(\lambda_a, \lambda_c) \) is non-negative and reflects the marginal effect of a change in \( \lambda_a \) on the norm targets’ estimation \( \hat{q}_a^* \) of the proportion of norm violators (with \( \hat{q}^* \) fixed) while the sign of \( \frac{\partial q^*}{\partial \lambda_a}(\lambda_a, \lambda_c) \) cannot be determined in general. Hence, \( q^*(\lambda_a) \) increases with \( \lambda_a \) only if the change in \( \hat{q}^* \) is negative and overcompensates the increase of \( \hat{q}^* \) by \( \lambda_c \). Thus, we expect that an increase in the norm targets’ suspiciousness causes a decrease of the proportion of norm violators in most cases.

Overall, we can predict that an increase in the norm targets’ or the inspector’s suspiciousness usually leads to a decrease in the equilibrium proportion of norm violators in the population. We provide a precise definition of “usually” whereas the definition with respect to a change in the inspector’s suspiciousness differs from that with respect to a change in the norm targets’ suspiciousness.

4.5 Robustness of the Results

So far, we assumed that all estimations \( \hat{q}_a, \hat{q}_c \) of the share of cheaters \( q \) satisfying Eq. (4.6) are feasible, i.e. restrictions are only caused by the share of caught cheaters \( \bar{q} \) which is public information and constitutes a lower bound for any estimation of \( q \). If we allow for a more complex estimation procedure, this may induce more restrictions. In case that both parties can use the knowledge of the utility functions and their past actions, the set of feasible estimations is further shortened. Let us assume that the agents and the inspector know that any estimation of \( q \) is restricted to the interval \([\bar{q}, 1]\). Both parties can use this knowledge to obtain boundaries for their effort and their counterpart’s effort at the previous time step: let \( e^*_a(\hat{q}), e^*_c(\hat{q}) \) denote the respective optimal effort obtained by plugging an estimation \( \hat{q} \) of the real share of cheaters \( q \) in Eq. (4.8) and Eq (4.7) alternatively. Then both parties can determine the minimum and maximum effort of the last period, and Eq. (4.5) leads to the restrictions

\[
p(e_a^*(1), e_c^*(\hat{q}(t)), \hat{q}(t)) \leq \frac{\hat{q}(t)}{\hat{q}_a(t)}, \ \frac{\hat{q}(t)}{\hat{q}_c(t)} \leq p(e_a^*(\hat{q}(t)), e_c^*(1), 1) \tag{4.17}
\]
as \( p \) is strictly increasing in the controllers’ effort and the share of cheaters and decreasing in the agents’ effort. Eqs. (4.6) and (4.17) lead to the condition

\[
q_i^-(t) = \frac{\bar{q}(t)}{p(e^*_a(\bar{q}(t)), e^*_c(1), 1)} \leq \hat{q}_i(t) \leq \min\left(\frac{\bar{q}(t)}{p(e^*_a(1), e^*_c(\bar{q}(t)), \bar{q}(t))}, 1\right) = q_i^+(t)
\]

for any estimator \( \hat{q}_i(t) \) of the real share of cheaters at time \( t \), \( i = a, c \). Note that \( \bar{q} \leq q_i^- \leq q \) always holds for \( i = a, c \). Additionally, one could allow for a memory of past actions. If at time \( t \) the agents and the inspector still have knowledge of their respective optimal efforts \( e^*_a(t-1) \) and \( e^*_c(t-1) \), the interval of feasible estimations can be shortened to

\[
q^-_a(t) = \frac{\bar{q}(t)}{p(e^*_a(\bar{q}(t)), e^*_c(t-1), 1)} \leq \hat{q}_a(t) \leq \min\left(\frac{\bar{q}(t)}{p(e^*_a(t-1), e^*_c(\bar{q}(t)), \bar{q}(t))}, 1\right) = q^+_a(t),
\]

\[
q^-_c(t) = \frac{\bar{q}(t)}{p(e^*_a(\bar{q}(t)), e^*_c(t-1), 1)} \leq \hat{q}_c(t) \leq \min\left(\frac{\bar{q}(t)}{p(e^*_a(1), e^*_c(t-1), \bar{q}(t))}, 1\right) = q^+_c(t).
\]

Note that in this case, the parties’ respective estimations \( \hat{q}_a \) and \( \hat{q}_c \) may differ as the knowledge of the respective effort at the previous time step is private. The interval of feasible estimations can be even more shortened if both parties know their counterpart’s estimation procedure. In this case, the agents could improve the lower and upper bounds for the inspector’s estimation (which are \( \bar{q} \) and 1 without this assumption) and vice versa.

With the above assumptions, the estimations \( \hat{q}_a \) and \( \hat{q}_c \) for the share of cheaters do not increase with the share of caught cheaters \( \bar{q} \) in general as \( q^+_a \) and \( q^+_c \) do not necessarily increase with \( \bar{q} \). Hence, Lemma 4.5(ii) does not hold in general, and Theorem 4.8 as well as Corollary 4.9 may not be valid for more complex estimation procedures. Nevertheless, the possibility of multiple equilibria is not affected by the above extensions of the basic estimation procedure.

### 4.6 Concluding Remarks

In this Chapter, we investigate how the confidence in a moral world affects normative behavior. We model the interaction between the targets and the inspectors of a social norm. We can show that so-called “dark figure” of undetected norm violations plays a pivotal role in understanding norm compliance and control. It is known from the “Thomas theorem” (Merton, 1995; Thomas and Thomas, 1928) that subjective beliefs (the moral confidence) substantially affect social reality. In the context of social norms, Popitz (1968) argues that the moral confidence of norm targets prevents norm violations. With respect to the inspectors, labeling theory (Becker, 1963; Lemert, 1967; Paternoster and Iovanni,
4.6. Concluding Remarks

1989) argues that the moral confidence of inspectors prevents inspections. We provide new insights by combining both approaches. We take the interaction between both mechanisms into account and reveal that the combination of both forces generates non-intuitive results. We consider the detection problem of norm violations in the context of public goods dilemmas. As in the context of tax evasion or doping in sports, the extent of norm violations is unknown to both the norm targets (e.g. taxpayers or athletes) and the inspectors. We propose a model that links the moral confidence with the actual rate of norm violations and control behavior. Here, the probability of detection is the result of an interplay between the norm targets and the inspectors. Each party’s behavior depends on the proportion of norm violators in the population which is unknown to both of them. Hence, the norm targets and the inspectors combine their respective (exogenous) moral confidence and public information about the proportion of detected norm violators to estimate the extent of norm violation in the population. We investigate how an equilibrium responds to a change in the norm targets’ and the inspector’s moral confidence.

It turns out that there is an “inverse” self-fulfilling prophecy effect: A small increase in the inspector’s moral confidence always leads to a lower real proportion of norm violators in equilibrium. Typically, we observe that an increase in the moral confidence of norm targets decreases the proportion of norm violators in equilibrium, although this may not hold in general. Our analysis shows that Popitz’ intuitive reasoning does not hold in general. The situation may change if there are external inspectors to detect norm violations: Our model considers inspectors who have an incentive to increase their control efforts when there are (or they believe that there are) many norm violations. Thus, the effect of moral decay on the side of the norm targets is overcompensated by the increased incentive of detection on the inspectors’ side.

In addition, we show that the dark field of norm violation can induce multiple equilibria of adherence to the cooperation norm. This is not possible for the case that the proportion of norm violators is public information. Hence, the equilibrium proportion of norm violators may not only depend on the parameters of our model but also on the initial proportion of norm violators in the population. As crime and control levels can stabilize at various fixed points, this result can be linked to the recent debate that punishment operates dissimilar in different societies (Herrmann et al., 2008), which might be due to distinct punishment cultures (Gintis, 2008).

A possible extension of our model could be addressed to the parameter $\lambda_a$ which represents a norm target’s moral confidence about the proportion of norm violators. We assume that this parameter is exogenous and constant. In reality, a taxpayer’s belief of the extent of tax evasion might be based on information about tax compliance in her network of friends.
Similarly, an athlete might have knowledge about drug use of teammates or athletes she regularly deals with. Hence, one could internalize $\lambda_a$ by introducing a network of norm targets: If we assume that a norm target can observe the actions of her neighbors, we can define $\lambda_a$, as the local proportion of norm violators in the neighborhood of norm target $i'$. Together with the proportion of detected norm violators $\tilde{q}$ (which is public information), this determines the norm targets’ estimation $\hat{q}_i$ of the real proportion of norm violators. As different local proportions of norm violators lead to different estimations, this increases heterogeneity in the system. An obvious question with respect to this setting is how the network topology affects a possible equilibrium proportion of norm violators.

Finally, we assume that the heterogeneous costs of our model are constant and in particular independent of the suspiciousness towards norm violation. However, norm targets might be less reluctant to violate the norm when they suspect a large extent of norm violation in the population. This behavior would be similar to Popitz’ original prediction, except that the perception of norm violation is not based on objective information but on subjective beliefs. This would counteract the inverse self-fulfilling prophecy effect with respect to the norm targets as an increase of the norm targets’ suspiciousness $\lambda_a$ would cause an increase in their heterogeneous costs and therefore more norm violations. With respect to an empirical validation, it is thus interesting to investigate whether there is a relation between the norm targets’ heterogeneous costs and their suspiciousness and if so, which of the two effects prevails.

In the following Chapter, we suggest and discuss an experimental design for the validation of our model assumptions and in particular the inverse self-fulfilling prophecy effect.
Chapter 5

The Effect of Nescience: An Experimental Design

In this chapter we suggest an experimental design to investigate to what extent the model assumptions and the theoretical predictions in the previous chapter are valid in the context of real interactions between norm targets and inspectors. After pointing out why experiments are more suitable compared to other methods of empirical validation, we specify an experimental design that extends a standard public goods game by incorporating inspections to detect possible violations of the contribution norm. Although the incentive structure of this setting generally corresponds to the norm targets’ and the inspector’s utility function in our model in Chapter 4, we reduced the model’s complexity by allowing only two possible values for the concealing effort of a norm violator and the control effort of the inspector. Further, we present the workflow of the experiment in all its steps. We suggest a measurement of the key variable “suspiciousness” that norm targets violate the norm and the assignment of subjects to experimental groups in order to test our hypothesis of an inverse self-fulfilling prophecy effect. After presenting the theoretical predictions of the experiment in detail, we finally outline a sketch of the statistical analysis to test our model assumptions and hypotheses.

5.1 Why Experiments?

What could be the appropriate method for validating our results? Data is usually available only for detected tax evasion, detected doping or detected crime in general. Nevertheless, there have been efforts to measure the dark field of crime. Most notably, there are perpetrator and victimization studies. Both reports, however, consist of very sensitive questions.
Perpetrators do not disclose information of their misconduct easily. Likewise, victims often feel embarrassed for revealing their personal sufferings and experiences of control loss. There are developments to improve the survey technology. Most notably, the randomized response and item count technique are methodologies which guarantee higher anonymity for the respondent. In the case of randomized response, respondents are asked to flip a coin and give a predetermined answer for tail (for example) and report the truth for head. In the case of item count, one part of the respondents receives a reduced list of uncritical statements and has to state the number of items, which apply for them. The other part of the respondents receives an extended list, which includes one added sensitive item. The difference of yes-statements between both groups estimates the rate of the sensitive behavior in the population. The underlying idea of both methodologies is to make it impossible for the researcher to know, which particular respondent has committed a crime or was a victim of crime. For an overview of measuring sensitive questions compare Tourangeau and Yan (2007), for an overview of measuring the dark field see Coleman and Moynihan (1996); Hindelang et al. (1981).

Survey data is unreliable as results often understate the degree of norm violation.\footnote{See Elffers et al. (1987) for an example with respect to tax evasion.} Further, our results provide an economic explanation why survey investigations may be inadequate. In terms of our model, a truthful report of norm violation by a norm target might lead to an increase of the norm targets’ and the inspectors’ moral confidence in the proportion of norm violators in the population. According to our results, this usually causes a higher detection probability and therefore reduces utility for norm violators. On the other hand, a norm target who adheres to the norm could profit from (anonymously) reporting norm violations by herself because a possibly increased moral confidence usually reduces the real proportion of norm violators and therefore the norm target’s disutility caused by norm violation by other norm targets. Thus, our model predicts incentives for untruthful reports for both, norm violators and norm compliant individuals. With respect to doping, one can identify an additional incentive for norm violators to conceal their fraud: If a survey indicates a high proportion of norm violators, some sponsors might reduce or cease their investment in the respective sport and thereby reduce all athletes’ income.

An experimental approach is more promising than survey investigations. Here, actual behavior is measured instead of potentially untruthful reports of behavior. Furthermore, decisions are associated with monetary incentives, which makes them more reliable and comparable to real norm violations in the field. As a baseline scenario, we suggest to use a public goods game. One type of subjects (the norm targets) can either contribute to a public good or violate the contribution norm. Another type of subjects (the inspectors)
can invest in detection of norm violators to gain rewards in case of success. Both types can invest different amounts of money to optimize concealment (norm targets) or detection (inspectors). To determine each subject’s ex ante moral confidence in the extent of norm violation, players and inspectors have to estimate the extent of norm violations among the norm targets (i.e. the dark field) before the game starts. Additionally, both types have to estimate the actual extent of norm violations after every period where the proportion of detected norm violators is made public. Herewith, one can investigate whether high confidence in norm compliance causes more norm violations. Furthermore, it would be possible to match norm targets with similar beliefs and examine whether these groups exhibit different proportions of norm violators on average. Finally, one could conduct experiments which vary the existence of a dark field.

5.2 A Simplified Public Goods Game With Unobservable Actions

As a baseline scenario, we suggest to use a public goods game. One type of subjects (the norm targets) can either contribute a fixed amount of money to a public good or violate the cooperation norm by contributing nothing. If norm violation is detected by the inspector (the other type of subjects), the respective norm target receives a sanction. Norm violators can invest in concealment to reduce their detection probability. Similarly, the inspector can invest in detection to gain rewards in case of successful detection of norm violators. Therefore, a strategy’s payoff generally corresponds to its respective utility defined by Eqs. (4.3) (norm targets) or (4.4) (inspector). However, we adjusted some details to the experimental environment: First, in contrast to our model where the concealment and the control effort are non-negative real numbers, we suggest to reduce the strategy space in order to keep the experiment as simple as possible. In the experiment, a norm violator’s concealing effort is either $e_a^{(1)}$ (low concealing effort) or $e_a^{(2)}$ (high concealing effort). Similarly, the inspector’s control effort is either $e_c^{(1)}$ (low control effort) or $e_c^{(2)}$ (high control effort). Additionally, we make the simplifying assumption that the detection probability does not depend on the proportion of norm violators in the population.\footnote{Note that this is not contradictory to the model assumptions as the discrete counterparts of (P6) and (P7) still hold.} The detection probability can then be defined by the $2 \times 2$ matrix $P = (p_{ij})$ with

\[
p(e_a^{(i)}, e_c^{(j)}, q) = p_{ij}, \quad i, j \in \{1, 2\},
\]
for any proportion $q$ of norm violators. To ensure that the experimental design is comparable to the model despite the reduction of complexity, we make the following assumptions:

$$(P1') p_{11} - p_{21} = p_{12} - p_{22}$$

$$(P3') p_{1j} < p_{2j}, \quad p_{i1} > p_{i2}$$

Here, (P1') is the discrete counterpart of the model assumption (P1) as the decrease of the detection probability caused by an increase of the concealment effort does not depend on the control effort and vice versa. (P3') ensures that the detection probability is decreasing in the concealment effort and increasing in the control effort and corresponds to the model assumption (P3). The assumption of decreasing marginal concealing and control effort in the model is obsolete in the experimental design as there are only two possible effort values. The remaining model assumptions (P2), (P4) and (P5) with respect to the detection probability are irrelevant in case of a discrete and finite space of concealing and control effort. The choice of the remaining model parameters (the amount of contribution $b_a$, the sanction cost $s$, the maximum detection benefit $b_c$ and the contribution multiplicator $r$) is in principle arbitrary but should guarantee sufficient variance in the subjects’ behaviors. For example, in case of exorbitant sanction costs $s$ of norm violation, one may expect that no norm target may violate the norm (for more details, see Section 5.3.2). Further, we assumed a continuum of norm targets in our model which of course can not be maintained in an experiment. In the latter context, the missing contribution of a norm violator has a (non-infinitesimal small) effect on her payoffs resulting from the public good. To summarize, a norm target’s expected payoff of a strategy $(d_i, e_{a_i}) \in \{0, 1\} \times \{e_{a_1}^{(1)}, e_{a_2}^{(2)}\}$ (norm adherence if $d_i = 0$, norm violation if $d_i = 1$) is

$$u_a(d_i, e_{a_i}) = d_i \left( b_a - sp(e_{a_i}, e_c) - e_{a_i} \right) - \frac{r}{n} \sum_{d_j = 0} b_a$$  \quad (5.1)$$

for given behaviors $d_j$ of the remaining norm targets and given control effort $e_c$ of the inspector. The heterogeneous costs $c_i$ are not considered for the norm targets’ payoff as they primarily represent the moral costs of norm violation. The inspector’s expected payoff of control effort $e_c \in \{e_{c_1}^{(1)}, e_{c_2}^{(2)}\}$ for given strategies of the norm targets is

$$u_c(e_c) = \frac{b_c}{n} \left( \sum_{d_i = 1} p(e_{a_i}, e_c) \right) - e_c.$$  \quad (5.2)$$

---

3As (P3') implies $p_{12} - p_{11} = p_{22} - p_{21}$, i.e. the increase of the detection probability caused by an increase of the control effort does not depend on the detection effort.

4Nevertheless, our experimental design allows us to estimate a norm target’s moral costs. For details, see Section 5.4.2.
5.2. A Simplified Public Goods Game With Unobservable Actions

At the beginning of the experiment, all participants are divided in groups of $G$ individuals with $G - 1$ norm targets and one inspector. In each of the $T$ rounds of the experiment, the participants of each group perform the actions given below. Before the experiment, all participants are informed about this procedure (including the payoff functions for norm targets and inspector and the fact that there are no false positive detections) so that the assumptions (K1)–(K5) of our model hold.

1. **Revelation of detected norm violations**: The number of detected norm violators in the previous round is published.

   *Note*: In the very first round of the experiment, steps 1 and 2 are omitted as there is no initial information about detection of norm violation.

2. **Estimating the overall extent of norm violation**: All participants estimate the number of norm violators in that round, and the participant with the best guess receives a reward. Further, all norm targets guess whether the inspector will choose high or low control effort in the current round whereas a correct guess will be rewarded.

   *Note*: Until the end of the experiment, the participants are not informed about the quality of their estimations.

3. **Endowment**: Each norm target receives $b_a + m_a$, the inspector $e_{c_2} + m_c$ units of money.

4. **Actions**: Each norm target chooses whether to add $b_a$ units of money to the public good or to violate the norm and contribute nothing. In case of norm violation, she chooses concealment effort $e_{a_1}$ or $e_{a_2}$. The inspector chooses control effort $e_{c_1}$ or $e_{c_2}$.

5. **Detection**: All norm violators are inspected, detection occurs randomly according to the detection probability resulting from the inspectors control effort and the respective norm violator’s concealment effort.

   *Note*: The identity of the detected norm violators is only revealed to the respective norm violator herself.

6. **Payment**: All participants receive payoffs according to Eqs. (5.1) (norm targets) or (5.2) (inspector).

   *Note*: In contrast to the inspector, no norm target is aware of her actual payoffs until the end of the experiment.
In each round, the participants thus have to choose their respective actions on the basis of the proportion of detected norm violators which constitutes the only objective information available. In our model, we assume that the agents and the inspector use this information to estimate the actual proportion of norm violators according to their subjective suspiciousness. These estimations are further utilized to determine the respective optimal actions. As we want to investigate if the participants use a similar mechanism, we need explicit information about their estimation of the actual extent of norm violation. Therefore, all participants have to estimate the proportion of norm violators after the proportion of detected norm violators is made public by the conductor of the experiment. As there is no intrinsic incentive for a correct estimation, the best guess is rewarded in each round. Additionally, the norm targets have to guess whether the inspector will choose high or low control effort in the current round. We can use this guess to measure to which extent the norm targets take the inspector’s control incentives into account with respect to their decision making (for details, see Section 5.4.2). We further require that the participants are not informed about the accuracy of their estimations until the end of the experiment. Therefore, we do not offer additional objective information about the real extent of norm violation. This setting corresponds to the situation for tax evasion, doping or criminal norms in general where we do not have objective information about the extent of norm violation except for the proportion or number of detected norm violators that is usually published by the inspection institution. For the same reason, the norm targets are not informed about their payoff at the end of a round of the experiment. If they were aware of this payoff, a norm target that adheres to the contribution norm could easily calculate the number of contributors by her share of the public good.

Further, we propose to inform a detected norm violator about her detection as this information usually occurs in the context of criminal norms.\footnote{Often, even all norm targets are informed about a detected norm violator’s identity, e.g. in case of doping.} However, this might induce memory effects in future rounds: A detected norm violator might be deterred from future norm violations by the negative experience of detection even if a cost-benefit-analysis favors a violation of the norm. Such a history dependent decision making is not accounted for in our model as only the overall extent of detection of norm violation is made public after each round and neither the norm targets nor the inspector are aware of which norm violator was actually detected. Nevertheless, we do not expect that a possible memory effect would interfere with the predicted inverse self-fulfilling prophecy effect. Hence, we prefer to keep the experiment close to the reality rather than adopting the mechanism of our model.

Finally, we suggest that the inspector receives a benefit $m_c$ in each round which is inde-
pendent of the success of detection in addition to the amount $e_{c_2}$ that can be invested to increase the detection probability. Therefore, we expect that the inspector is less reluctant towards a risky investment in her control effort as she has a guaranteed income in each round by $m_c$. For the same reason, the norm targets receive a fixed amount of $m_a$ units of money in each round that they cannot contribute to the public good.

5.3 Analysis of the Experimental Design

The experimental design is tailor-made with respect to the assumptions of our model in Chapter 4. Similar to the analysis in Section 4.3 and Section 4.4, we can derive the dynamics of the experiment in case the participants maximize their expected payoff according to Eqs. (5.1) (norm targets) and (5.2) (inspector). As the expected payoffs are very similar to the expected utility for the norm targets and the inspector in our model (see Eq. (4.3) and Eq. (4.4) respectively), many of our model’s theoretical properties still hold in the experimental setting. However, we will show in this section how our modifications and in particular the simplifying assumption of binary concealment and control efforts affect the dynamics. Using these theoretical predictions, we formulate “reasonable” conditions for the parameters of the experiment and discuss a possible configuration. In contrast to the previous analysis which is restricted to mean value dynamics, we hereby additionally use multi agent simulations to investigate how non-deterministic effects (in particular with respect to the detection probability) could affect the experiment.

5.3.1 Properties of the Theoretical Dynamics

First, we determine the optimal concealment and control efforts $e_{a}^{*}$ and $e_{c}^{*}$ for a given (estimated) proportion of norm violators. Without loss of generality we hereby assume that choosing the respective low effort results in zero cost, i.e. $e_{a_1} = e_{c_1} = 0$. Then, a norm violator that maximizes her expected payoff prefers the high concealment effort $e_{a_2}$ if and only if

$$s > \frac{e_{a_2}}{p_{11} - p_{21}}$$

(5.3)

whereas this preference is independent of the control effort as we required (P3’). This implies in particular that the optimal concealing effort does not depend on the estimated proportion of norm violators. For a given estimated proportion of norm violators $\hat{q}_c$, the
inspector prefers the high control effort $e_{c_2}$ if and only if

$$b_c \hat{q}_c > \frac{e_{c_2}}{p_{12} - p_{11}}. \quad (5.4)$$

Again, this preference is independent of the concealment effort by (P3'). Hence, the optimal control effort (non-strictly) increases with the estimated proportion of norm violators. More precisely, there is a threshold $q^*_c > 0$ where the inspector prefers the high control effort if $\hat{q}_c > q^*_c$ and the low control effort if $\hat{q}_c < q^*_c$. In case that $q^*_c > 1$, the inspector prefers the low control effort independent of the estimated proportion of norm violators. Further, given an estimated proportion $\hat{q}_a$ of norm violators, a norm target that maximizes her expected payoff (without considering moral costs of norm violation) prefers to violate the norm if and only if

$$b_a > \frac{n}{n - r} s p(e^*_a, e^*_c(\hat{q}_a)) \quad (5.5)$$

whereas $e^*_c(\hat{q}_a)$ denotes the estimated optimal control effort of the inspector based on the norm target’s estimation of the proportion of norm violators and $n$ denotes the number of norm targets. Different to the context of the model where we assumed a continuum of norm targets, a norm target additionally has to account for missing public good benefit $\frac{z}{n} b_a$ resulting from her own contribution in case of norm violation. If $\hat{q}_a > q^*_c$, a norm target anticipates that the inspector chooses the high control effort while she assumes a low control effort if $\hat{q}_a < q^*_c$.

In case the norm targets consider heterogeneous costs $c_i$ distributed according to a function $F$ as assumed in our model, this implies that the reduction of the space of possible concealment and control efforts strongly simplifies the shape of the iteration function $g_{\tilde{q}}$ as defined in Eq. (4.13). Here the proportion of detected norm violators $\tilde{q}$ at time $t + 1$ is determined by the respective proportion at time $t$, i.e. $\tilde{q}(t + 1) = g_{\tilde{q}}(\tilde{q}(t); \lambda_a, \lambda_c)$. According to our assumptions on the estimation procedure of the norm targets and the inspector in Eq. (4.12) which state that an estimation of the real proportion of norm violators is determined by the proportion of detected norm violators and the norm targets’ (the inspector’s) suspiciousness $\lambda_a (\lambda_c)$, we can reformulate

$$\hat{q}_i > q^*_c \iff \tilde{q} > \frac{q^*_c - \lambda_i}{1 - \lambda_i}, \quad \hat{q}_i < q^*_c \iff \tilde{q} < \frac{q^*_c - \lambda_i}{1 - \lambda_i}$$

for $i = a, c$. Therefore, with respect to the experimental design, we obtain

$$\tilde{q}(t + 1) = \begin{cases} q_{11} = F\left((1 - \frac{z}{n}) b_a - s p(e^*_a, e_{c_1}) - e^*_a p(e^*_a, e_{c_1})\right) & \text{if } \tilde{q}(t) < \frac{q^*_c - \lambda_a}{1 - \lambda_a}, \tilde{q}(t) < \frac{q^*_c - \lambda_c}{1 - \lambda_c} \\ q_{21} = F\left((1 - \frac{z}{n}) b_a - s p(e^*_a, e_{c_2}) - e^*_a p(e^*_a, e_{c_1})\right) & \text{if } \tilde{q}(t) > \frac{q^*_c - \lambda_a}{1 - \lambda_a}, \tilde{q}(t) < \frac{q^*_c - \lambda_c}{1 - \lambda_c} \\ q_{12} = F\left((1 - \frac{z}{n}) b_a - s p(e^*_a, e_{c_1}) - e^*_a p(e^*_a, e_{c_2})\right) & \text{if } \tilde{q}(t) < \frac{q^*_c - \lambda_a}{1 - \lambda_a}, \tilde{q}(t) > \frac{q^*_c - \lambda_c}{1 - \lambda_c} \\ q_{22} = F\left((1 - \frac{z}{n}) b_a - s p(e^*_a, e_{c_2}) - e^*_a p(e^*_a, e_{c_2})\right) & \text{if } \tilde{q}(t) > \frac{q^*_c - \lambda_a}{1 - \lambda_a}, \tilde{q}(t) > \frac{q^*_c - \lambda_c}{1 - \lambda_c} \end{cases}$$
for a given proportion of detected norm violators $\tilde{q}(t)$. Hence, $g_{\tilde{q}}$ is piecewise constant. Different to our model in Section 4 which exhibits continuous concealment and control efforts, the simplifying assumption of binary efforts does not guarantee the existence of an equilibrium as the iteration function $g_{\tilde{q}}$ might be discontinuous. If $\lambda_a \geq \lambda_c$ ($\lambda_a \leq \lambda_c$), the condition for $q_{12}$ ($q_{21}$) cannot be fulfilled. Hence, $g_{\tilde{q}}$ does not exhibit more than three possible proportions of detected cheaters and at most two if the norm targets and the inspector exhibit identical suspiciousness. According to our assumptions, we have $q_{11} \geq q_{21}$ since the share of norm violators decreases if the norm targets anticipate an increased control effort.\footnote{If $F(b_a - s p(e_a, e_c) - e_a^* - c_i) = F(b_a - s p(e_a, e_c) - e_a^* - c_i)$, i.e. if there are no norm targets with appropriately heterogeneous costs to be detained from norm violation by the anticipated high control effort, the proportion of norm violators remains unchanged.} This also leads to a decrease in the proportion of detected norm violations as the detection probability remains unchanged as the inspector’s actual control effort remains the same. Similarly, we have $q_{22} > q_{21}$ as an increased detection probability will increase the proportion of detected norm violators if the proportion of norm violators remains constant. Analogous statements hold for $q_{12}$, and we summarize

$$q_{11} \geq q_{21} < q_{22}, \quad q_{11} < q_{12} \geq q_{22}. $$

The relation between $q_{11}$ and $q_{22}$ cannot be determined in general and depends on the model parameters including the distribution of the heterogeneous costs.

### 5.3.2 The Choice of Parameters

We now use the theoretical analysis of the experimental design to determine suitable parameter configurations in the sense that non-trivial and diversified behavior of the participants can be observed. A reasonable condition to such a configuration is that for low control effort, norm violation should pay off for a norm target with high concealing effort, i.e.

$$\left(1 - \frac{r}{n}\right) b_a > s p_{21} - e_{a_2} $$

while the moral costs of norm violation are not considered. Similarly, norm violation should not pay off for low concealment effort and high inspection effort, leading to

$$\left(1 - \frac{r}{n}\right) b_a < s p_{12}. $$

In case of identical levels of concealment and control effort, we suggest that norm violation is still beneficial compared to norm adherence, i.e.

$$\left(1 - \frac{r}{n}\right) b_a > s p_{11}, s p_{22}. $$
However, the payoff difference should be smaller than that in case of superior concealment effort. The reason is that we want to emphasize the economic incentive to violate the norm in order to maintain the experiment's character as a public goods game and a social dilemma. Additionally, norm targets usually account for moral costs in case of norm violation and therefore tend to prefer norm adherence in case of payoff-indifference between the two behavioral patterns. Therefore, we want to avoid that norm violation is too unattractive ex ante. With respect to the sanction cost $s$ for a detected norm violator, we first require that it exceeds the benefit of violating the contribution norm, i.e. $s > b_a$. The sanction mechanism would probably not be perceived as a serious punishment by the norm targets if the costs of norm violation were not at least as high as the contribution to the public good. Hence, there would be few economic incentives for the norm targets not to violate the norm, and we would probably observe very few norm adherent participants independent of the control behavior and in little interaction between the norm targets and the inspector during the experiment. On the other hand, the higher the sanction costs for norm violation, the more norm targets will be deterred from norm violation irrespective of the anticipated quality of controls. Therefore, we will probably also observe little interaction between the norm targets and the inspector if these costs are too high. Further, a low proportion of norm violators implies an even lower proportion of detected norm violators. This is especially problematic if the number of norm targets taking part in the experiment is low. In this case, we would observe only little variation in the proportion of detected norm violators which constitutes the only objective information for participants with respect to the extent of norm violation in the population. For example, if only none or one norm violator is detected in each round of the experiment, we cannot expect much feedback between this quantity and participant behavior. We expect similar difficulties with respect to few interaction effects in case of very high or very low detection probabilities. Hence, we propose that a “medium” value of $s$ is appropriate for the experiment, for example

$$s = 1.5 b_a.$$ (5.9)

Further, we expect that in case of norm violation, the norm targets would prefer the high concealment effort even if the expected benefit was equal to norm violation with low concealment effort. The reason is the lower detection probability in case of high effort so that it is less likely for a norm violator to incur the moral costs of norm violation captured by the heterogeneous costs $c_i$ in our model. Hence, we propose to go against this preference.

\[^7\text{More precisely, the benefit of norm violation is } (1 - \frac{1}{R}) b_a < b_a \text{ as there is a finite number of norm targets.}\]
and give an incentive for low concealment effort by

$$sp_{11} < sp_{21} + e_{a_2}. \quad (5.10)$$

Hence, a norm violator that maximizes her expected payoff always prefers the low concealment effort independent of the estimated extent of norm violation in the population. Finally, we suggest that the costs for high detection effort account for half of the contribution to the public good, i.e.

$$e_{a_2} = 0.5 \cdot b_a. \quad (5.11)$$

In this case, the difference in expenditures between norm violation with high and low concealment effort equals the difference between adhering to the contribution norm and norm violation with high concealment effort.

<table>
<thead>
<tr>
<th>parameter</th>
<th>$b_a$</th>
<th>$s$</th>
<th>$b_c$</th>
<th>$e_{a_2}$</th>
<th>$e_{c_2}$</th>
<th>$r$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>3</td>
<td>4.5</td>
<td>12</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>(0.4 0.7)</td>
</tr>
</tbody>
</table>

Table 5.1: The suggested parameter configuration satisfying Eqs. (5.6)-(5.11). This configuration should be checked in a pretest of the experiment.

According to these considerations, we suggest the parameter configuration depicted in Table 5.1. We assume that the number of norm targets is $n = 10$. One can easily verify that this setting fulfills the conditions in Eqs. (5.6)-(5.11). In Figure 5.1 we depict an agent based simulation of the experiment. The dynamics are not deterministic even if participants maximize their expected payoff due to the stochasticity of the detection of norm violators. We observe that for the above parameters, the participants in fact respond to a change in the proportion of detected norm violators that constitutes the only objective information. If this proportion exceeds the threshold induced by Eq. (5.4), it pays off for the inspector to choose the high control effort. As this is anticipated by the norm targets, norm violation yields less payoff than adherence even in absence of moral costs of norm violation according to Eq. (5.5). Hence, all norm targets contribute to the public good so that the proportion of detected norm violators becomes zero. In the next round, the inspector therefore prefers the low control effort resulting in a positive proportion of norm violators as norm violation pays off for sufficiently low moral costs according to Eq. (5.5).

One can further observe the inverse self-fulfilling prophecy effect as we increase the norm targets' and the inspector's suspiciousness. In this case, the iteration function $g_q$ shows

---

We assume that the available infrastructure allows this value of $n$. As the benefit of norm violation is $10 \cdot \frac{r}{n}$, the number of norm targets affects the conditions for the parameters.
Figure 5.1: Simulation of the experiment with \( n = 10 \) norm targets (including the theoretical iteration function \( \tilde{g}_q \)) and parameters according to Table (5.1) for different levels of suspiciousness with heterogeneous costs equally spaced in the interval \([0, b_a]\). The dynamics for the proportion of norm violators \( q \) and detected norm violators \( \tilde{q} \) show the inverse self-fulfilling prophecy effect as we increase the norm targets’ and the inspector’s suspiciousness from 0.05 in (a) to 0.35 in (d). Here, the dashed line indicates the threshold \( \tilde{q}^\# \) with respect to the control effort: For \( \tilde{q} > \tilde{q}^\# \), it pays off for the inspector to exert high control effort while she chooses low control effort otherwise. This is anticipated by the norm targets who prefer the low concealment effort independent of \( \tilde{q} \). As norm violation does not pay off in case of high control effort, all norm targets adhere to the norm in the following round when \( \tilde{q} \) is above the threshold. For low suspiciousness ((a) and (b)), there is a stable fixed point of the theoretical dynamics of \( \tilde{q} \). Due to the random detections of norm violators, the behavior of norm targets is not constant.
5.3. Analysis of the Experimental Design

Figure 5.1: (continued) As we increase the suspiciousness, the fixed point is destabilized in (c) leading to oscillations between a nonzero proportion of norm violators paired with low control effort and complete norm adherence paired with high control effort. Due to the further increase of the suspiciousness in (d), high control effort pays off for the inspector even if no norm violation is detected. As this is anticipated by the norm targets, complete norm adherence is a stable equilibrium. Overall, the average proportion of norm violators decreases with the suspiciousness which confirms the inverse self-fulfilling prophecy effect.

that the threshold in the proportion of detected norm violators inducing complete norm adherence is decreased. For a suspiciousness of $\lambda_a = \lambda_c = 0.25$, this decrease destabilizes the system in the sense that there is no stable equilibrium anymore due to the discontinuity of the iteration function. Hence, one can observe oscillations in the participants behavior: As soon as there are norm violators, the inspector increases the control effort which is anticipated by the norm targets and leads to complete norm adherence. Therefore, the inspector decreases the control effort leading again to a positive fraction of norm violators. If some norm violators are detected, the procedure recurs. Overall, this leads to a decreased
average proportion of norm violators which constitutes the inverse self-fulfilling prophecy effect. If we further increase the norm targets’ and the inspector’s suspiciousness, the threshold proportion of detected norm violators finally goes to zero. In this case, all norm targets adhere to the contribution norm and the inspector chooses the low control effort at any timestep.

5.4 Statistical Analysis

In this section we describe the analysis of the experimental data in order to test the inverse self-fulfilling prophecy effect and the assumptions of our model. Here, we cannot avoid an imbalance with respect to the size of data for the norm targets’ behavior on the one hand and the inspector’s behavior on the other hand as each cycle of the experiment contains $G - 1$ norm targets and only one inspector. Nevertheless, we aim for a sufficient statistical significance by conducting $G$ cycles of the experiment while rotating the role of the inspector among the participants.

5.4.1 The Inverse Self-Fulfilling Prophecy Effect

To test our hypothesis of an inverse self-fulfilling prophecy effect, we propose to group the participants by their suspiciousness before the experiment. Our hypothesis is validated if we observe that those groups with a high suspiciousness exhibit a low proportion of norm violators. We can obtain a proxy measure of the norm targets’ and inspector’s suspiciousness $\lambda_a$ and $\lambda_c$ as follows: Before the group assignment and before the first round of the experiment (and thus without information about the proportion of detected cheaters), all participants have to guess how many norm targets will violate the norm on average during the experiment.\(^9\) We can use this estimation $\hat{\lambda}_i$ of participant $i$ to determine her initial confidence in the contribution norm and thereby obtain a measure for the suspiciousness parameters $\lambda_a$ and $\lambda_c$ of our model. The participants can then be divided in two groups $A$ and $B$ where each member of $A$ has a higher value of the proxy $\hat{\lambda}$ than any member of group $B$. According to our hypothesis, group $A$ should exhibit a lower average proportion of norm violators than group $B$ in the context of our experiment. Should we observe the predicted effect, we will further investigate whether this result was also caused by a correlation between the norm targets’ suspiciousness and their moral costs of norm violation. According to our model, these costs are constant and not affected by an increase the norm targets’ suspiciousness. However, it is possible that there is a

\(^9\)Again, the best estimation will be rewarded.
connection between moral costs and suspiciousness. To capture a norm target’s moral costs of norm violation, we introduce

\[ \hat{c}_i(t) = \begin{cases} b(t) & \text{if } b(t)(d_i(t) - \frac{1}{2}) < 0 \\ 0 & \text{if } b(t)(d_i(t) - \frac{1}{2}) \geq 0 \end{cases} \]

with

\[ b(t) = \begin{cases} (1 - \frac{r}{n})b_a - sp(e_{a_i}(t), \hat{e}_c(t)) - e_{a_i}^* & \text{if } d_i(t) = 0 \\ (1 - \frac{r}{n})b_a - sp(e_{a_i}(t), \hat{e}_c(t)) - e_{a_i}(t) & \text{if } d_i(t) = 1 \end{cases} \]

representing norm target \( i \)'s expected benefit of norm violation at time \( t \) based on her estimation \( \hat{e}_c(t) \) of the inspector’s control effort. In contrast to our model, \( \hat{c}_i(t) \) can be negative in the context of the experiment if a norm target violates the norm although adherence would yield a higher expected payoff according to her estimation. A norm target’s estimated moral costs is the average

\[ \hat{c}_i = \frac{1}{T} \sum_{t=1}^{T} \hat{c}_i(t) \]  

(5.12)

over all \( T \) rounds within a cycle of the experiment. Further, we can use Eq. (4.12) to define a participant’s estimated average suspiciousness

\[ \hat{\lambda}(t) = \frac{\hat{q}(t) - \tilde{q}(t)}{1 - \tilde{q}(t)} \]

with \( \hat{q}(t) \) denoting the participant’s estimation of the proportion of norm violators and \( \tilde{q}(t) \) denoting the proportion of detected norm violators at time \( t \). The classification of the cycles can then be conducted by the respective inspector’s average estimated suspiciousness

\[ \hat{\lambda}_c = \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}_c(t). \]  

(5.13)

If we observe a positive correlation between a norm target’s moral cost \( \hat{c}_i \) and her estimated suspiciousness \( \hat{\lambda}_a \), this should reinforce a possible inverse self-fulfilling prophecy effect while we expect the opposite in case of a negative correlation.

A disadvantage of the above procedure is that the number of norm targets within one experiment is halved by the partition of the participants. As discussed before in Section 5.3.2, this may inhibit the interaction between the norm targets and the inspector. Further, the ex-ante partition is only meaningful if the proxy \( \hat{\lambda} \) is adequate and in particular if the suspiciousness does not vary during the experiment. Both disadvantages can be avoided
if we restrict ourselves to a test of the inverse self-fulfilling prophecy effect with respect to the inspector, i.e. if we only vary the inspector’s suspiciousness $\lambda_c$ and investigate its effect on the extent of norm violation in the population. In this case, no partitioning of the participants is necessary as we already change the role of the inspector between different cycles of the experiment with the same individuals. Comparing two cycles, all norm targets except one are identical so that the average suspiciousness should be approximately identical while the inspector’s suspiciousness might be significantly different. Further, the classification of the cycles according to the respective inspector can be conducted ex-post and without additional information from the data obtained in the regular experiment. Additionally, the theoretical evidence for an inverse self-fulfilling prophecy effect with respect to the inspector’s suspiciousness is stronger than the effect with respect to the norm targets’ suspiciousness as discussed in Section 4.4.3. Hence, it might be more likely to observe the effect in the experiment if we stick to a variation of the inspector’s suspiciousness.

5.4.2 Testing the Model Assumptions

Besides a test of our hypothesis of an inverse self-fulfilling prophecy effect, the experiment allows us a detailed validation of our assumptions concerning the norm targets’ and the inspector’s decision making and their respective estimation of the extent of norm violation in the population. In our model, we assume that the norm targets and the inspector estimate the proportion of norm violators at each timestep by a simple mechanism that combines objective information (the proportion of detected norm violators) with subjective beliefs (the suspiciousness). Based on this estimation, they determine their respective optimal strategy while incorporating the knowledge about their counterpart’s incentives. By evaluating the experimental data, we are able to test to what extent the participants’ decision making resembles this mechanism.

First, we can check the model assumptions with respect to the estimation procedure: (E1) and (E2) reflect that an increase in the proportion of detected norm violators or the suspiciousness increases the estimation of the proportion of norm violators. These assumptions could be identified with a positive correlation between the estimation and the suspiciousness or the proportion of detected norm violators. Here, we estimate the participants’ respective suspiciousness by $\hat{\lambda}$ according to Eq. (5.13). Further, the norm targets’ respective guess with respect to the inspector’s control effort in each round of the experiment can be utilized to investigate whether they conduct a cost-benefit-analysis as in our model. We expect a positive correlation between the estimated proportion of norm
violators and the estimated control effort of the inspector. Further, the estimated control effort should be negatively correlated with the binary variable $d_i$ that indicates whether norm target $i$ violates the norm ($d_i = 1$) or not ($d_i = 0$).

With respect to the homogeneity of the suspiciousness (E4), one can determine to what extent the estimated suspiciousness varies over time and among the participants. In the latter case, heterogeneity should be reduced by a grouping of the participants to test the inverse self-fulfilling prophecy effect (cf. Section 5.4.1). Further, a participant’s suspiciousness can generally be assumed to be private information (E3). Even if we group the participants by their suspiciousness, they are not aware of the relevance of the suspiciousness for the partitioning or the partitioning procedure itself.

Finally, our model predicts that the optimal concealment effort of a norm violator is independent of the endogenous model parameters and thus constant whereas the inspector’s optimal control effort varies with the estimated proportion of norm violators. Hence, we expect a higher variance of the control effort as compared to the concealment effort.

### 5.5 Concluding Remarks

One could vary the experiment for additional research questions. For example, we can inform the participants about the proportion of norm violators after each round in order to investigate how the average proportion of norm violation is affected compared to the original setting with a darkfigure of norm violation. Here, our model does not offer a general prediction in this case as the equilibrium proportion of norm violators in a darkfigure setting can be higher or lower than without a darkfigure depending on the norm targets’ and the inspector’s suspiciousness.

In our model in Chapter 4, we assume that the norm targets’ heterogeneous costs are independent of their suspiciousness towards norm violation. However, as already discussed in Section 4.6, the participants’ reluctance with respect to norm violation may decrease if their suspiciousness increases. This would result in more norm violations which counteracts the predicted inverse self-fulfilling prophecy effect with respect to the norm targets. Hence, we might observe that in an experiment according to our design, the group with the higher estimated suspiciousness exhibits more norm violations compared to the other group. However, a relation between the norm targets’ heterogeneous costs and their suspiciousness should not affect the inverse self-fulfilling prophecy effect with respect to the inspector, i.e. the extent of norm violation should decrease if the inspector’s suspiciousness increases.
Chapter 6

Conclusion

This thesis investigated the emergence and enforcement of social norms from an interdisciplinary point of view. Hereby, we followed the tradition of mathematical sociology (Coleman, 1964; Edling, 2002; Fararo, 1973; Leik and Meeker, 1975) by combining sociological and economic theory with mathematical modeling and analysis as well as computer simulations. Our main results include a general game theoretic framework for social influence in models of opinion dynamics which we can formally identify as a coordination process (Chapter 2). We further extended the opinion dynamics model of Deffuant et al. (2000) to investigate the emergence of local cultures in clusters of firms (Chapter 3). We showed that this mechanism fosters consensual business practice of firms which constitutes a requirement to realize the benefits of their co-location. With respect to cooperation norms, we investigated the influence of subjective beliefs about norm adherence on normative behavior and identified an inverse self-fulfilling prophecy effect which is antipodal to the hypothesis by Popitz (1968) (Chapter 4). We then provided an experimental design for an empirical validation of this effect (Chapter 5). In the following, we discuss our approach, suggest extensions of our research and dwell in particular on our model assumptions with respect to the degree of the individuals’ rationality and heterogeneity. Further, we comment on the validation and relevance of our results.

With the exception of the local cultures model, our research is not restricted to specific instances of norms. Instead, the general models and results refer to a certain class of norms. For example, the model in Chapter 4 is applicable to cooperation norms where detection of norm violators is costly and performed by an external inspection institution. Further, our results are qualitative in the sense that we derived general trends that are mostly independent of the model parameters. An example is the inverse self-fulfilling prophecy effect: The exact extent of the decrease of norm violation after an increase of the suspiciousness towards norm violation depends on the model parameters and possibly
also the initial state of the system. The general effect is however independent of the parameters.

A common basis for our models is a rational choice approach as we assumed that individuals maximize their respective utility. The models comprise a microscopic definition of the interactions between individuals whose actions determine the emergence or enforcement of a norm on the macro level. We thereby presumed that individuals are boundedly rational: Their strategic horizon is restricted to the actual time step, and they cannot foresee how other individuals will respond to their behavior. Therefore, the degree of their rationality is between the perfect rationality of standard game theory and evolutionary game theory where individuals are not forward-looking at all. However, our mechanism of myopic best response to previous behavior is not the only implementation of bounded rationality. For example, Young (1993) allows agents to make mistakes with a certain probability and to only partly observe the behavior of other agents. Therefore, their degree of rationality and knowledge is reduced. On the other hand, they exhibit a variable size of memory while in our models, agents can only observe behaviors at the previous time step. Hence, future research could investigate whether our results are robust with respect to different implementations of bounded rationality.

Further, we allowed for heterogeneous individuals: They have different initial behavior or differ in their reluctance to violate the norm. However, the heterogeneity in our models is limited and can be extended in future research. In the local cultures model, all firms have identical open-mindedness. As a possible extension, one could vary open-mindedness (akin to Lorenz (2008)) and investigate the influence on the likelihood of consensus. With respect to our model in Chapter 4, norm targets are assumed to act homogeneously except for individual moral costs of norm violation. In particular, their suspiciousness towards norm violation is identical so that all estimations by norm targets are identical. In real life, this suspiciousness may be influenced by an individual’s personal experience. For example, a norm target might be able to observe norm violations conducted in her circle of family or friends. Then, her suspiciousness towards norm violations in the population will certainly be influenced by these observations. In future research, one could introduce a network between the norm targets and endogenize a norm target’s suspiciousness by the local frequency of norm violation in her neighborhood and then investigate how the network topology affects the extent of norm violation.

Besides its function to check the consistency of verbal theories, the main purpose of formal modeling is to generate testable hypotheses (Carley, 1997). With respect to our model of local cultures in clusters of firms, we show by means of computer simulations that our model with its evolving interaction network is conducive to consensual behavior among a
majority of firms in a cluster (compared to the model of Deffuant et al. (2000)). A next step in advancing the model would consist in a benchmark against data. Unfortunately, the key model parameter (the firms’ open-mindedness with respect to interaction with different firms) is very difficult to measure. As a result, any data investigation would probably have to rely on qualitative, case-study evidence of how firms choose to interact with each other and whether a concern for one’s past partners does exist. Such findings would give an inclination of whether the mechanisms proxied in the model are actually at work. Beyond a data benchmark the link between open-mindedness and group effects on consensus could be investigated experimentally. Participants could be surveyed on open-mindedness and would be allocated to two groups accordingly. The experiment could then study in how far the consensus dynamics differ between both groups.

In order to validate the model in Chapter 4 and in particular the inverse self-fulfilling prophecy effect, we pointed out that laboratory experiments constitute an appropriate methodology. We suggested an experimental design based on a simplified version of the public goods game in our model. Further, we discussed which parameter configuration and which measures are suitable for a validation. However, the simplified game is still more complex than an average laboratory experiment. Therefore, except for an implementation of the experimental design, future research could also validate the claim of Popitz (1968) that nescience about norm violations enhances norm adherence when the detection problem is neglected. Although this problem is certainly relevant, a neglect reduces the complexity of this approach. First, one could develop a theoretical model of how individuals react to information about the (global or local) extent of norm violation where threshold models similar to those of Granovetter (1978) might be a starting point. Here, in contrast to our model, detection of norm violators would not be modeled explicitly. Then, Popitz’ prediction and the model could be validated in a laboratory experiment which would be simpler compared to our suggested design as it does not include detections of norm violation.

The relevance of our model of social influence builds on its generality. It provides a formal basis for diverse models of opinion dynamics (e.g. Axelrod, 1997; Deffuant et al., 2000; Hegselmann and Krause, 2002) that essentially existed in parallel so far. Additionally, the formal identification of the general model as a coordination process shows that opinion dynamics can be related to coordination norms. In particular, its results with respect to the emergence of consensual behavior may contribute to the research on the emergence of this type of norm, e.g. the emergence of local cultures. With respect to our results on the enforcement of cooperation norms, we can interpret the recent development of doping in sports against the background of our model and the inverse self-fulfilling prophecy effect: While in the decades before approximately 1990 very few athletes were detected for drug
use, there is strong evidence that doping was performed area-wide in some countries, in particular in the German Democratic Republic (Franke and Berendonk, 1997). Although at the beginning of this century there are (also relatively) more detections of drug using athletes (e.g. Brissonneau and Depiesse, 2006, p. 164), the proportion of drug users is assumedly lower (e.g. Leonard, 2001). Hence, although an increased number of detections could be interpreted as a hint for more violations of the anti-doping norm, the opposite is supposed, also due to the higher effectiveness of controls. In the context of our model, this development can be explained by an increase in the suspiciousness towards doping, possibly due to intensified media coverage.

Finally, it shall be mentioned that the strict distinction of research questions for coordination norms on the one hand and cooperation norms on the other hand is not universally valid. For example, norm enforcement is not completely irrelevant with respect to coordination norms: Although there is no incentive to drive on the left side of the road in general, some people nevertheless occasionally do so and are fined for this. Similarly, there may be more than one solution for a cooperation problem. If conflicts arise when individuals act on the basis of distinct solutions, they have to coordinate towards one norm. However, the distinction is useful if we consider the general incentive structure of norm relevant situations.


Flache, A.; Macy, M. (2006). Why more contact may increase cultural polarization. In: *Session mathematical sociology I at 101st annual meeting of the american sociological association, Montreal, 14 August 2006*.


Curriculum Vitae

Personal Details

Name Patrick Christian Groeber
D.O.B. 23 June 1978
Nationality German
Address Werdstrasse 104, 8004 Zurich, Switzerland
Phone (office) +41 (0)44 632 8432
Phone (home) +41 (0)43 539 7741
Email pgroeber@ethz.ch

Education

Since 04.2006 Ph.D. student at the Chair of Systems Design, Department of Management, Technology and Economics, ETH Zurich, Switzerland
10.2000–12.2005 Studies in mathematics, University of Bremen, Germany
• Masters degree in Mathematics (Diplom Mathematiker)
• Faculty prize for best diploma thesis of the year
• Thesis title: “Convex combination of discrete dynamic processes using the example of Parrondo’s paradox”
09.1997–02.2000 Apprenticeship as assistant mathematical technician at STN Atlas Elektronik, Bremen, Germany
• Examination at the CCI Cologne, Germany
1997 Abitur (German GCE A-levels equivalent), Delmenhorst, Germany
Employment

since 04.2006 Research assistant at the Chair of Systems Design, ETH Zurich, Switzerland

03.2000–08.2000 Assistant mathematical technician at STN Atlas Elektronik, Bremen, Germany

09.1997–02.2000 Apprentice as assistant mathematical technician at STN Atlas Elektronik, Bremen, Germany

Publications

peer reviewed:


(basis for Section 2.4)

(basis for Chapter 3)

working paper:

(basis for Chapter 4)

(basis for Chapter 2)