MRes/PhD in Political Science

Norms and Games: Realistic Moral Theory and the Dynamic Analysis of Cooperation

Kai Spiekermann

London School of Economics and Political Science

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

To Lydia, Klaus, and Robin.

Declaration

I certify that the thesis presented by me in 2008 for examination for the MRes/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others, and that the extent of any work carried out jointly by myself and any other person is clearly identified in it.

Abstract

The thesis investigates how social norms are enforced. It consists of two parts. The first part establishes the concept of "realistic constraints for moral theory" based on the "ought implies can principle". Different notions of feasibility lead to different degrees of moral realism. Game theory and computational modelling are the appropriate instruments to determine feasibility constraints for realistic moral theory. They allow for a dynamic perspective on norm enforcement, in contrast to more static approaches. The thesis discusses the use of computational models and game theory from a philosophy-of-science point of view. I conclude that computational models and game theory can inform moral theory if they are understood as sources of realistic constraints.

The second part uses two agent-based models to explain the enforcement of social norms. In the first model, agents play one-shot, two-person prisoner's dilemmas. Before the game, agents have a better than random chance to predict which strategy the others are going to play. Cooperative agents do well if they are able to pool their information on the strategies of others and exclude defectors. The second model analyses repeated multi-person prisoner's dilemmas with anonymous contributions. The players are situated in a social space represented by a graph. Agents can influence with whom they are going to play in future rounds by severing ties. Cooperative agents do well because they are able to change the interaction network structure.

I conclude by connecting the findings with debates in moral philosophy and evolutionary theory. The results obtained have implications not only for the emergence of cooperation and social norms literature, but also for theories of altruism, research on social network formation, and recent inquiries by behavioural economists into the effects of group identity.

Contents

List of Figures

List of Tables

Introduction

1. Introduction

The conduct of humans is regulated by social norms. No society could work without a set of prescriptions as to how individuals should deal with each other. Some of these norms are institutionally enshrined, written down as laws, officially policed, and their transgression punished. But most social norms cannot be found in statute books, no police officer watches over their fulfilment, and no court would impose a penalty on you for violating social norms. Nevertheless, social norms must still be enforced somehow, otherwise they would cease to be norms. This thesis discusses the subtle mechanisms to enforce social norms. Understanding these subtle mechanisms, I will argue, enables us to understand a crucial element of human sociality.

The argument in this thesis has two parts: a theory part and a modelling part. The theory part can be distinguished further into a moral theory component and a methods component. The moral theory component in part I (chapter 3) is concerned with the development of realistic moral theory. I argue that moral theory should not only be interested in what is just, but also in what is feasible. Utopian norms cannot be complied with in a non-perfect world. To identify realistic moral norms, we need a definition of feasibility. A suitable definition of feasibility must take the dynamic processes of norm enforcement into account. Therefore, I argue, agent-based models are the appropriate tool to conduct this feasibility analysis. The methods component in part I (chapters 4 and 5) discusses methodological issues regarding computational modelling and explanations of norm emergence. I propose a philosophy-of-science analysis of computational modelling in the social sciences. In particular, I suggest that social science models should be highly idealised models rather than high fidelity models. I also address recent trends to use evolutionary theories for computational models and for the analysis of social norms. The modelling part of this thesis (chapters 6 and 7) consists of two agent-based models to explore a mechanism used by groups to enforce cooperation: social exclusion. The first model assumes that agents are able to exclude defectors *ex ante* in one-shot multi-person prisoner's dilemmas if agents are slightly better than random in identifying defectors and if they pool their information. The second model shows that defector exclusion can also work in multi-person games with-

out *ex ante* information if the game is played on a network and agents can influence the network structure. These two models demonstrate that exclusion, and the threat of exclusion, are powerful mechanisms to secure cooperation.

The two parts are linked. Part I introduces a notion of feasibility that is best researched by applying agent-based models. It also debates the philosophy-of-science foundations for these agent-based models. This leads to the modelling part, where I present two concrete agent-based models which can be interpreted in light of the framework offered in the first part.

Taken together, the two parts of my argument answer two questions:

1. How do groups enforce social norms?

2. Which norms can emerge under which conditions?

My answer to the first question is that exclusion is a central mechanism of norm enforcement. I answer the second question by analysing feasibility constraints for norms. Only feasible norms can emerge, and a norm is feasible if it can be enforced in such a way that compliant agents do not lose out.

Some readers may not agree with all steps of my argument. One could agree with each of the single chapters, but not with the connections between them, or one could reject ideas from some chapters, while supporting others. For example, a reader might reject my analysis of social exclusion in chapters 6 and 7, but still agree with my argument for realistic moral theory in chapter 3 (or the other way round). A different reader might disagree with my methodological commitments regarding model-building, but still concur with my ideas on realistic moral theory. Less likely is a rejection of my methodology, while agreeing with my agent-based models, because the models rest on my methodological arguments. In any case, the chapters of this thesis can be read independently. I hope, of course, that readers agree with all of them, and also with the overall argument holding them together.

1.1. Social Dilemmas and Norms

How do social norms link with social dilemmas of cooperation? To answer this question, I have to define social dilemmas of cooperation, and show how social norms are connected to these dilemmas.

Cooperation poses a social dilemma when it is collectively beneficial if all individuals cooperate, but even more beneficial for each single individual not to cooperate

("defect"). This leads to a dilemma: If all individuals are utility-maximisers, they all defect. The dilemma is that each individual ends up worse-off compared to the situation where all cooperate. If the individuals could only decide between "all cooperate" and "all defect", they would choose "all cooperate". But since they each make an individual choice, cooperation fails.

The classic example is the prisoner's dilemma. Assume that players Row and Column face monetary payoffs as stated in table 1.1. The first number in each cell is the payoff for Row, the second the payoff for Column. Suppose for the sake of argument that Row and Column only care for the maximisation of monetary payoff (we will come back to this dubious assumption). For both players the best outcome arises if they defect while the other player cooperates. The second-best outcome is mutual cooperation, the third-best mutual defection, the worst is being the "sucker", that is to cooperate while the other player defects.

		Column	
		cooperate	defect
Row	cooperate	$2, 2$	$-1, 3$
	defect	$3, -1$	$0, 0$

Table 1.1.: A prisoner's dilemma.

With these preferences over outcomes, both players should see that defection is the strictly dominant strategy. Since Row cannot influence the choice of Column (and vice versa), Row realises that she gets a better outcome if she defects (rather than cooperates), given that Column cooperates. She also gets a better outcome when she defects, given that Column defects. Row makes the same considerations because the payoffs are symmetrical. Payoff-maximising agents defect in a prisoner's dilemma. Defection/defection is the only strict Nash equilibrium, i. e. both players play their strictly best response, given the strategy the other player plays. Row and Column are not happy with the outcome. They would have preferred mutual cooperation over mutual defection. The dilemma in the prisoner's dilemma is that players do not get what they could get if cooperation was possible. This 2-person dilemma can be extended to multi-person dilemmas. The qualitative problem stays the same: Collectively, cooperation is

preferable, but individually, defection is the strictly dominant strategy and all rational players defect.

Social norms can be part of the solution for social dilemmas. However, they might also pose a social dilemma themselves when it comes to the question of compliance. Assume that a norm prescribes cooperation in a prisoner's dilemma. The social dilemma is now a norm compliance problem: Both agents see the collective benefit from mutual compliance, but they are both tempted to violate the norm. If both players obey the norm, they both cooperate. The norm transforms the prisoner's dilemma into a cooperation game, if the norm is successfully enforced. More technically speaking, the norm changes the payoffs in the game. This transformation can only work if non-compliance is costly. The costs of violating the norm must be higher than the expected benefit from it. Therefore, the enforcement of norms and the sanctioning mechanism used are crucial to understand whether and how a social norm solves a social dilemma.

Not all social dilemmas are prisoner's dilemmas. Consider the stag hunt game. Table 1.2 gives the payoffs for Row and Column. We still assume for simplicity that agents are

		Column	
		stag	hare
Row	stag	2, 2	0, 1
	hare	1, 0	1, 1

Table 1.2.: A stag hunt.

payoff maximisers. Row and Column can either hunt stag or they can hunt hare. They will only be able to catch stag if they do it together. Hares can be caught individually (Skyrms, 2004). There are two pure strategy Nash equilibria, one where both hunt stag together, and one where both hunt hare individually. Clearly, it would be best to hunt stag together. The problem is that the decision to hunt stag comes with a risk: If the other player walks off to hunt hare, the lonely stag hunter does not catch anything.

Whether the two players coordinate to hunt stag depends on their risk aversion and their beliefs what the other player is going to do. Risk averse players will hunt hare unless they are almost certain that the other player hunts stag. In cases like these, a social norm can solve the dilemma in two ways: Either the norm creates expectations

that all (or most) players hunt stag. Or the social norm changes the preferences such that the stag hunt turns into a cooperation game. In the first case, the players succeed in coordinating on a stag hunt because they expect each other to hunt stag. In the second case, the two players hunt stag together because their preferences have changed.

These examples show that I look at social norms from their functional side. More specifically, I am interested in those social norms that have the function to solve social dilemmas. For these norms I use the term "social norm of cooperation", which I define as follows:

Definition: Social Norm of cooperation. A *social norm of cooperation* is a norm that prescribes actions which, when the norm is obeyed by all participating parties, produce greater average payoffs, compared to outcomes where the norm is not in place and agents pursue individual payoff maximisation.

Social norms of cooperation exist because they help to produce cooperation gains in settings that are potential dilemmas of cooperation. Social norms are not enforced by the state and its institutions, but by less formal, subtle means. The costs of transgressing a social norm are therefore often not of a pecuniary nature, or at least not directly. Rather, the costs are often psychological: Norm violators are excluded, they lose social status, or are publicly shamed. In this thesis I focus on exclusion and its effectiveness as an informal sanction.

Not all social norms can work equally well, given the nature of humans and the dilemmas of cooperation and coordination they encounter. A norm can only be sustained as a meaningful norm if it is enforceable. I use the concept "enforcement" in a broad sense: It is a social arrangement that transforms individual preferences such that a social dilemma dissolves. Some norms are more difficult to enforce than others. If free-riding is easy, sanctioning hard, and detection of cheaters difficult or expensive, then norms are difficult to enforce. Some norms are unenforceable. I argue that unenforceable norms will not survive as norms because free-riding frustrates the compliant agents and leads to a breakdown of the norm. This implies that actually existing norms are enforced somehow. Since humans often cooperate in situations that seem to be social dilemmas, one can ask how these norms are enforced, even if conventional forms of sanctioning (such as policing and fines) do not work. My answer is that most social norms rely on subtle sanctioning mechanisms, particularly the threat of social exclusion. Understanding subtle forms of sanctioning helps us to understand which social norms are enforceable and therefore realistic, and which social norms are unenforcable and therefore utopian.

But what about intrinsic motivations to comply with a norm? Couldn't there be a moral motivation to obey a social norm, even if the enforcement of the norm fails? The answer is: Yes, it is perfectly possible that people are motivated by genuine moral motivations, independent from any enforcement mechanism. I do not question that moral motivations play a role. The problem in social interactions is that morally motivated agents often meet agents driven by less noble motivations. In a social dilemma, even a majority of morally motivated agents might fail to sustain cooperation when they are challenged by "selfish" free-riders who constantly do better than the morally motivated agents. Therefore, my argument for realistic norms does not rule out moral motivations, it only points out that a norm can be robust only if it can be enforced so as to keep the more "selfish" individuals in check.

Many social norms are so deeply embedded in our social life that we hardly notice them, until they are violated. While writing this introduction I am sitting in the LSE library. Around me are about 20 other students. It is reasonably quiet. No one is talking. Everyone is dressed. No one is digging their nose. I claim that this behaviour is prescribed by social norms. There are probably people around who would, if they were to consider only their narrow individual preferences, prefer to violate one or more of these norms. The norm against talking is particularly tempting to be violated in this situation. Everybody knows that there is no effective institutional sanction in place to stop people from talking (one could call library security, but the effort is great and nothing will come from it). Nevertheless, no one talks. Why? Because violating this norm will induce psychological costs. The other library users will make sure that talkers realise how annoying and socially unacceptable their behaviour is.

Not all social norms are that trivial. Consider the problem of providing public security. The norms against killing, rape, robbery, etc. are not only moral, but also legal norms and are enforced by the police and other appropriate institutions. However, it would be a mistake to believe that these norms are exclusively enforced by institutional means. That this is not the case can be seen in anarchical, failed states. Once anarchy reigns, even robust attempts to enforce legal norms with police and military intervention often fail to reinstate civility. In failed states, ordinary policing is insufficient to enforce even the most basic norms. So why does norm enforcement work in functioning states? It works because legal enforcement is backed up with social enforcement. Legal norms are usually social norms as well. They are not only obeyed because people fear to be caught by the police, but because they fear the social punishment from their peers. Nothing short of an Orwellian state could do enough policing to enforce all the norms we obey without questioning every day. If social norms and their informal sanctioning

mechanisms break down, a society is in deep trouble. Thus, understanding social norms and their enforcement will enable us to understand the fundamental glue that keeps societies together.

It seems that the analysis of social norms can be conducted from two fundamentally different perspectives. On the one hand, normative political philosophers ask what makes a social norm *just*. Their stance is entirely normative, and rarely touches upon the positive, empirical questions of how social norms are actually enforced. On the other hand, legal scholars and economists, as well as analytical political scientists and sociologists, are engaged in an entirely positive analysis of social norms and their enforcement. Surely, it is important to distinguish between descriptive and prescriptive analysis, between "what is" and "what ought to be". However, this distinction has led to an unhealthy isolation of normative theory from positive analysis. What can be prescribed must—at least to a certain extent—depend on what is feasible. If the discussion of what ought to be is completely separate from what can be done, the result will be utopian theory. But if normative theory wants to engage with the world as it is, one must take feasibility constraints into account. Therefore, the normative analysis must be embedded in a positive analysis of what is feasible.

Regarding social norms, the issue of feasibility must be looked at from different angles. Whether it is feasibile to comply with a norm depends on physical facts about the world, psychological facts about the subject, and the social dynamics of the environment the subject is situated in. The last point is the most interesting. I argue that what is feasible for one subject does not only depend on what the subject can physically do and what is compatible with the subject's psychological state. It also depends on other people's actions. For norm compliance, it is not only important whether this subject is physically and psychologically able to comply with a norm, it also depends on how many other people comply. Norm compliance can be costly, and the benefit from compliance depends on how many other people comply. If the rate of defection becomes too high, compliance is often no longer feasible for a rational individual.

To analyse the social feasibility constraints I use agent-based models. Agent-based models are the appropriate tool to capture the dynamic aspect of social interactions. They are ideal to explore suspected social mechanisms that are currently not well understood by carrying out simple "if–then" experiments of social dynamics on a computer. Naturally, agent-based models are not a *panacea*. They come with their own methodological problems, in particular regarding the relation between model and target system, and the challenge of model validation. Nonetheless, I argue that agent-based models promote the research into the social feasibility constraints of norm compliance

and cooperation.

In the remainder of this introduction, I give an overview of the argument. The thesis consists of two parts: In the first part, I show how game theory and computational modelling can advance our understanding of norm enforcement and of realistic moral theory. In the second, I develop formal models of subtle norm enforcement mechanisms and utilise computater simulations to explore their behaviour.

1.2. Social Norm Enforcement, Modelling, and Evolution

The thesis begins with a brief literature review in chapter 2. After that, chapter 3 sets up a framework to discuss the realism of social norms, in particular moral norms. While most normative theory is only concerned with what is just, I ask what is feasible. Starting from a discussion of the "ought implies can" principle, I argue that a positive analysis of feasibility should complement the normative debate. Impossible actions cannot be demanded from agents. To develop a realistic approach to moral theory, I introduce the concepts of feasibility, moral realism, and moral utopianism by applying a possible worlds semantic.

What is feasible for agents depends on physical restrictions, but also on psychological conditions and the dynamics of the social environment the agents operate in. Therefore, to determine which actions are feasible, it is necessary to analyse psychological constraints and constraints stemming from the dynamic interactions with other agents. I argue that agent-based models are ideal to conduct feasibility analyses because they are well suited to capture the dynamic element of social interactions.

Public goods dilemmas are a suitable example for feasibility constraints. Such dilemmas occur when the provision of a public good depends on the contribution of several people. The problem is that each individual prefers to let others contribute, while the individual "free-rides". Since all individuals prefer free-riding over contribution, the provision of the public good fails, unless contributions are motivated externally. Suppose there is a norm prescribing cooperation in a specific type of public goods dilemma. How feasible is this norm? Naturally, this depends on agent-specific factors, for instance whether the agent has the means to contribute, on the agent's character, and so on. But it also depends on the behaviour of other agents. It is easier to contribute to a joint project when everyone else is contributing too. It is harder when few people contribute. Should a norm prescribe actions that force an agent to be exploited constantly, it becomes increasingly unfeasible for this agent to obey this norm, as his resources dwindle, while other agents thrive. Thus, the feasibility of contributing in a

public goods dilemma depends on how others behave.

Chapter 4 adresses methodological questions that arise when computational models are used in the social sciences and philosophy. First, I look into the problem of representation: In virtue of what does a model represent the world or aspects of the world (Frigg, 2006)? Many philosophers of science think that the key to answering this question lies in the alleged similarity relation between model and target system. The problem is that no proper concept of this similarity relation is forthcoming for all the different types of models used in the sciences. Maybe a unified account of this similarity relation should not be expected. For the highly idealised models used in the social sciences, the dissimilarities are often more apparent than the similarities. Social scientists deliberately distort and omit aspects of reality to keep their models simple. Idealisation is crucial for computational models in the social sciences. Since our abilities to model complex social systems are limited and our knowledge is poor, we are better off restricting the analysis to simple *social mechanisms*. Social mechanisms give an account of potential underlying causal relations between inputs and outputs by deliberately omitting other intervening factors or mechanisms. The role of mechanisms for explanation has been debated in the philosophy-of-science literature (Machamer, Darden and Craver, 2000). For the social sciences, mechanisms are a useful level of description because they focus on the behavioural dispositions of complex systems. Even if we do not understand all causal processes in a complex system, we often know how the system is disposed to react to certain inputs. This is a typical research situation for the social scientist. In these cases it makes sense to describe the social system as a social mechanism with certain disposition. The aim of the researcher is to understand the mechanism's behaviour, but also to describe the mechanism's internal operations to account for this behaviour (Glennan, 2002).

To approach the analysis of social mechanisms, the modeller has to apply two processes of idealisation: isolation and generalisation. Isolation is necessary to (artificially) separate the proposed social mechanism from its complex interactions with other processes. The researcher operates with very strong and usually counterfactual *ceteris paribus* assumptions. The second process, generalisation, is necessary to show that the hypothesised social mechanism applies in many different situations, i. e. that it is a robust mechanism. Using isolation and generalisation means that the models make many unrealistic assumptions. Nevertheless, I argue that the model is still similar to the target system insofar as model and target system instantiate some similar properties. Which properties need to be instantiated depends on the aims of the modeller. The appropriate properties are those that render the model relevant. However, the

analysis of relevance regarding model-building is difficult, and may have to operate on a case-by-case basis (Teller, 2001). Simple, highly idealised models are often preferable because they are easier to understand, analyse, and validate. In addition, for simple models the similarity between model and target system is often easier to show than for more complex ones.

In the second part of chapter 4 I distinguish models along two dimensions: highly idealised versus high fidelity models, and thought experiments versus empirically grounded models. I argue for highly idealised models. An empirical foundation is desirable, but pure thought experiments can be useful as well, especially in the exploratory phase of theory development. High fidelity models rest on methodologically shaky grounds because they are usually underdetermined by the available data and are likely to produce artefacts.

Chapter 5 addresses some issues arising from recent attempts to explain the emergence of norms and moral systems with evolutionary theory. This opens a wide field, which cannot be scrutinised in depth in this thesis. I restrict myself to comment on some theories that have immediate relevance for my own project. After a brief introduction into biological and cultural evolution, I discuss the options we have to explain norms from an evolutionary angle. Cultural evolution seems to be of greater relevance because the amazing development of different norm systems in the last few thousand years has been moving too fast to be explained by biological evolution. However, the predictive power of cultural evolution is questionable. In chapter 5 I espouse a skeptical position. Norms may be "new replicators" (Dennett, 2006), but it is unclear what determines the fitness of a norm. If no consistent definition of norm fitness can be given, the predictive power of an evolutionary theory of norms is low.

Instead of putting too much hope into the predictive power of evolutionary theories of norms, I suggest to change the whole approach. Rather than trying to explain how a process of evolution (be it biological or cultural) has produced human norms systems as we know them, I use game theoretical considerations and simulations to derive possibility results for social norms. By "possibility results for norms" I mean arguments or evidence that the social norm in question can be consistently applied by an agent or agents without doing too badly.

Sociobiologists often claim that the analysis of human evolution (both biologically and culturally) allows us to draw conclusions about behavioural patterns of humans today. For example, many sociobiologists believe that human norms of fairness are the result of evolutionary processes, because having certain norms of fairness led to an increase in adaptive fitness (compared to many other conceivable norms of fairness that

did not survive the evolutionary process). There are many problems with these kinds of arguments. I do not want to suggest that they always fail—to the contrary, some results in evolutionary psychology are very promising—but I think that most of these arguments overplay their hand.

What I intend to do is more modest, but, I will argue, not less interesting. Rather than using evolutionary theory to derive direct implications about the content of human behavioural patterns (and ultimately about human norms and morality), I argue that evolutionary game theory should be used to infer *side constraints for social norms.* Firstly, I argue that norms can make prescriptions that neither maximise payoffs nor evolutionary fitness. Many norms are not made to maximise anything. So I argue against a direct effect of evolution on the prescriptive content of social norms. Secondly, however, there are certain prescriptive contents that realistic norms cannot have. If an agent follows a norm that exposes him to constant and severe exploitation, it is likely that the norm collapses: Norms that force you to be the "sucker" are unlikely to see high levels of compliance. A norm can only be socially stable if its followers do well enough, compared to other people who do not obey it.

1.3. Formal Computational Models of Norm Enforcement

To determine how norms of cooperation are enforced and sanctioned, it is important to understand dynamic social processes. Until recently, most models of cooperation assumed unstructured environments where agents meet randomly to interact with each other. Take Axelrod's (1984) "Evolution of Cooperation". In his famous computer tournaments, each participating program had to play repeated prisoner's dilemmas against all other competitors.[1] It is well known that TIT-FOR-TAT emerged as the winner in Axelrod's tournament. TIT-FOR-TAT cooperates in the first round and then simply copies the move the other player made in the previous round. While it is now clear that TIT-FOR-TAT is not an optimal strategy[2], it often does well because it is conditionally "nice".

[1] There is an important theoretical difference between finitely and infinitely repeated prisoner's dilemmas. If finite, no cooperation should emerge because both agents use backward induction: Both agents know that there is no point cooperating in the last round. Since both defect in the last round, it makes no sense to cooperate in the second last round, etc. However, if a game is infinite or has a sufficiently high probability to be repeated, no such backward induction can be applied. For these latter infinitely repeated 2-person prisoner's dilemmas, the folk theorem states that any outcome can be an equilibrium. In Axelrod's tournament, a finite game was implemented. But no strategy used the finite character of the game, therefore the results would also apply to an infinite setting.

[2] See the literature review below for references.

Axelrod's results were pathbreaking, not only for its methodology (computer simulations), but also for popularising the problem of cooperation and the modelling tools to approach it. The repetition of games is one important factor to explain the emergence of cooperation. When games are repeated, players have a "shadow of the future". While cooperation never pays in the one-shot prisoner's dilemma, it can pay in an infinitely repeated setting, because agents can reciprocate behaviour and punish each other for defection. Being cooperative, as long as the opponent is cooperative, pays.

In recent years, the analysis of cooperation has turned to other mechanisms that had long been neglected. Repeated play is certainly one important factor, but it is by far not the only one. Other important mechanisms include indirect reciprocity, spatial structures, evolutionary dynamics, and cues to distinguish cooperators from defectors. Empirical advances in behavioural economics and social psychology teach us that processes of exclusion and group identity can be crucial for obtaining high levels of cooperation. In addition, Axelrod's "Evolution of Cooperation" only dealt with 2-person games, but most realistic social dilemmas have multi-person characteristics.

Chapter 6 analyses a simple process of defector detection and group formation to explain high levels of cooperation in n-person prisoner's dilemmas. The argument combines three ideas: translucency, the epistemic superiority of groups according to the Condorcet Jury Theorem, and group assortation. Translucency is a concept used by David Gauthier (1986). In Gauthier's argument, agents have different dispositions as to how to play in a situation with the payoffs of a prisoner's dilemma. For example, an "unconstrained maximiser" always defects, while an "unconditional cooperator" always cooperates and a "constrained maximiser" cooperates when he beliefes that he plays against an unconditional cooperator or a constrained maximiser. An agent is "translucent" if other agents can (fallibly) recognise the agent's disposition.[3] The idea that agents are slightly better than random in predicting cooperation and defection of others in a prisoner's dilemma is empirically supported (Frank, Gilovich and Regan, 1993; Brosig, 2002). Assume that agents can refrain from playing if they believe that the opponent will defect. With such a setting, translucency helps to improve the performance of cooperating agents. However, it would be surprising if the prediction of a single agent was good enough to do better than defectors. It is likely that people are somewhat better than random at predicting the dispositions of other agents. It is not likely that the forecast is good enough to give cooperators an advantage.

This problem can be solved when agents do not act alone, but are able to pool their translucency information. The insight that groups can make better judgements

[3] An agent would be "transparent" if this recognition was not fallible.

than individuals, given certain conditions, goes back to the Marquis de Condorcet (see Estlund et al., 1989; List and Goodin, 2001). In my model, I use the insights from the Condorcet Jury Theorem and combine these with a process of group formation. The idea is that agents go through a period of assortation (Sethi and Somanathan, 2003): One group is able to restrict admission of additional agents by voting on the candidates willing to join the group. All members of the group are interested in admitting only cooperators, because the members will have to play an n-person prisoner's dilemma with all agents of their group, once the group formation process has finished. If the group is sufficiently large, and if all members of the group have independent translucency information about the candidate willing to join the group, then the group as a whole should be able to distinguish cooperators from defectors by pooling their information. It is likely that the group succeeds to form a cluster of like-minded cooperators, which allows cooperators to do well, even though they play a game with the payoff structure of an n-person prisoner's dilemma.

Using computer simulations, I show that the predicted results do indeed obtain over a wide range of parameters. Cooperators do well because they pool their information and cluster. This model demonstrates how even simple assumptions about social structure increase the chances for cooperators to harvest cooperation gains. Exclusion is a powerful and cheap punishment mechanism, and probably one of the most important sanctions in human interaction. The results are remarkable because cooperation can be maintained, even though the game played is an anonymous n-person prisoner's dilemma in terms of payoff, where cooperation is particularly difficult to achieve.

Chapter 7 takes this analysis further. Social structure, defined as the environment for interaction, is now modeled as a non-directed graph. A vertex (or node) represents an agent. Edges connect vertices. If two agents (vertices) are connected by an edge, it means that the two agents interact with each other, that is, they are involved in a game with each other. I assume that the graph is not complete, meaning that not all possible edges exist in the graph. This means that agents interact with some, but not all other agents.

The core idea of the model is to let the network structure emerge endogenously. Connected agents play n-person prisoner's dilemmas with each other and, depending on the outcomes of the games, the agents have the option to delete connections to agents if they wish to do so. Deleted edges are replaced by new edges between two randomly chosen vertices. The graph changes over time. The number of vertices and edges, however, remains constant.

Although this model may sound abstract, it represents a concrete and important phe-

nomenon: Individuals can influence their social environment. If previous interactions have frustrated an agent's ambitions, the agent can "move on" and stop interacting with those players that caused the frustration. Humans maintain "successful" social contacts, but they avoid people who have exploited them. Again, the mechanism of exclusion proves to be a powerful one to maintain cooperation in hostile environments. My computer simulations show that cooperators can do well in this structured environment, even though the detection and exclusion of defectors is quite difficult.

Chapter 7 extends the analysis of chapter 6. But there is one important difference: The translucency-assortation model in chapter 6 plays anonymous games which are technically one-shot games, because the choices of agents are not influenced by future games. The network model in chapter 7, by contrast, involves a repeated game. Agents restructure the network in response to previous outcomes.

The thesis ends with a conclusion, linking the results of Parts I and II. Part I offers a framework for the analysis of social norms. It also discusses methodological questions regarding model building and evolutionary explanations of norms. Part II builds on these theoretical foundations and models processes of norm enforcement with dynamic agent-based models.

Acknowledgements

This thesis would not have been possible without the intellectual and personal support from my supervisor Christian List. He always managed to challenge and encourage me at the same time, and I have learned much more than could possibly be written down in this thesis. I am very grateful for all the time he spent with me over the last four years. Cécile Fabre was an excellent advisor to me whenever I asked her for help and support. Katrin Flikschuh listened and gave good advice when I was unsure about the right directions for my PhD.

During the last four years I received feedback from many people. I want to mention the members of the LSE Choice Group, the Political Theory Workshop and the Political Science and Political Economy Workshop, the latter two at the Department of Government. Special thanks go to my friend Gabi, who had to listen to many complaints and worries, as well as to my friend Andreas, with whom I spent many hours discussing the state of our theses in particular, and the state of the world in general.

I am grateful to the German National Academic Foundation for generous financial and intellectual support. The Philosophy Program at the Research School of Social Sciences, Australian National University, invited me in both 2006 and 2007. Crucial parts

of my thesis were written in Canberra, and I am grateful for the feedback and many excellent discussions I enjoyed there. Special thanks to Geoff Brennan, Bob Goodin, Kim Sterelny, Lina Eriksson and John Matthewson. In June 2007 I participated in the Santa Fe Institute Complex Systems Summer School. I want to thank the Santa Fe Institute and the National Science Foundation (grant 0200500) for this excellent programme.

My partner David has always been there to cheer me up and give me new strength. I doubt that I would have finished the thesis without him.

Finally, I dedicate the thesis to my family, Lydia, Klaus, and Robin.

2. Literature Review

The breadth and depth of the literature on the emergence and maintenance of cooperation and norm compliance in dilemma situations is impressive. A complete review is beyond the scope of the thesis. One fascinating feature of the field is its genuine interdisciplinarity. Mathematicians, game theorists, biologists, political scientists, and sociologists have all contributed to it. This forces me to restrict myself to some core findings with further references in the cited literature.

2.1. Games and Preferences

Before I review the literature that is directly related to my field of research, I have to introduce the notion of a game and clarify my position regarding the concept of preferences.

Game theorists use slightly different notations and terms to describe games. But apart from notational differences, there is consensus on the analysis of simultaneously played games, which are usually called strategic games (see for example Osborne and Rubinstein, 1994; Binmore, 1992). I distinguish between game form and game. The game form of a strategic game consists of a set of players, a set of strategies for each player, and outcomes.[1] The outcomes are strategy profiles (Aumann, 1989, p. 8; Osborne, 2004, p. 13). Consider the prisoner's dilemma (PD) from the introduction (table 1.1). There are two players (I have called them Row and Column, but usually they are just reprented by numbers, i. e. 1 and 2). Each player has two strategies available: cooperate or defect. The set of outcomes is the set of strategy profiles: {(cooperate, cooperate), (cooperate, defect), (defect, cooperate), (defect, defect)}, where the first element of each tuple denotes player 1's strategy, the second 2's strategy. It is also possible to attach payoffs to the outcomes. In table 1.1 in the previous chapter we have payoffs for each player for each strategy profile. For example, the outcome (defect, defect) has the payoffs (0,0) for the two players.

[1] Terminological difficulties can arise because strategies are sometimes called actions. The terms actions and strategies can be used interchangeably for strategic games, but they come apart for games with more than one stage, as I explain below.

The game form determines the rules of the game, but it does not contain assumptions which strategies the players are going to play. To move from the game form to the game, we need the players' preferences over outcomes. If we assume that players care for maximising their payoffs in the example of table 1.1, then we can derive a preference order over the set of strategy profiles for each player. Let $\succ$ signify the preference relation "is strictly preferred over". Then player 1 (Row) has the preferences (defect, cooperate)$\succ$(cooperate, cooperate)$\succ$(defect, defect)$\succ$(cooperate, defect). Analogously, player 2 (Column) has the preferences (cooperate, defect)$\succ$(cooperate, cooperate)$\succ$(defect, defect)$\succ$(defect, cooperate). To summarise, a game form consists of players, strategies for each player, and outcomes. A game is established by a game form and the players' preferences over the outcomes.

The most important solution concept for games is the Nash equilibrium. Binmore (2007, p. 18) explains it succinctly for the case of two players: "A pair of strategies is a *Nash equilibrium* in a game if and only if each strategy is a best reply to the other." In the PD, the only Nash equilibrium is (defect, defect). Other games can have more than one Nash equilibrium.

While discussing the PD, we moved from payoffs to preferences by assuming that people are payoff-maximisers. Sometimes this assumption is taken for granted, but it is far from clear that players are always payoff-maximisers. Definining preferences clearly is particularly important when analysing social dilemma situations where agents seem to make choices against their interests. There is disagreement as to how preferences should be interpreted. Many game theorists prefer to take preferences as ordinal preference rankings. They also assume that preferences are immediately action-guiding, i. e. if you strictly prefer strategy D over C, you will do D. Since preferences cannot be observed, but actions can, this usually leads to the theory of revealed preferences: The choices of agents reveal their preferences. In the example, if an agent consistently chooses D over C, one assumes that the agent prefers D over C.

When we accept these assumptions, defection is what rational players do in the PD *by definition.* For example, Ken Binmore (2006), calling his two players Alice and Bob, and talking of "dove" and "hawk" rather than "cooperate" and "defect", insists that

> "[i]n the Prisoners' Dilemma, to write a larger payoff for Alice in the bottom-left cell of the payoff table than in the top-left cell therefore means that Alice would choose *hawk* in the one-person decision problem in which she knew in advance that Bob had chosen *dove.* Similarly, writing a larger payoff in the bottom-right cell means that Alice would choose *hawk* in the

> one-person decision problem in which she knew in advance that Bob had chosen *hawk*. The very definition of the game therefore says that *hawk* is Alice's best reply when she knows that Bob's choice is *dove*, and also when she knows his choice is *hawk*. So she does not need to know anything about Bob's actual choice to know her best reply to it. It is rational for her to play *hawk* whatever he is planning to do." (Binmore, 2004, p. 12)

Amartya Sen argues for a more complex view on preferences. He distinguishes between sympathy and commitment, arguing that commitment "drives a wedge between personal choice and personal welfare" (Sen, 1977). In "Rational Fools" he rejects the view that agents always choose what increases their own welfare. In the case of commitment, welfare and choice come apart: When committed, agents may act against their welfare because they feel a duty to do so. In later papers, Sen explicitly introduced different notions of preference: "self-centered welfare", "self-welfare goal", and "self-goal choice" (Sen, 2002, p. 206–224). Daniel Hausman, largely sympathetic to Sen's critique, nevertheless proposes to use only one definition of preferences. As I will follow his proposal in my thesis, I cite him at length:

> "An agent's preferences consist of his or her overall evaluation of the objects over which preferences are defined. This evaluation implies a ranking of these objects with respect to everything that matters to the agent: desirability, social norms, moral principles, habits—everything relevant to evaluation. Preferences thus imply *all-things-considered rankings.*" (Hausman, 2005, p. 37, my emphasis)
>
> "Sen is right to maintain that 'A theory of human behaviour—even on economic matters—calls for much more structure and many more distinctions.' But it doesn't follow that it needs multiple notions of *preference.* On the contrary, it seems to me that Sen's concern to draw distinctions can be accomodated at least as well by distinguishing sharply *between* preference (as all-things-considered ranking) and other things. Rather than taking expected advantage to be one concept of preference, one can distinguish *between* preference and expected advantage. Instead of taking one concept of preference to refer to 'mental satisfactions', one can distinguish between 'mental satisfactions' and the extent to which preferences are satisfied." (Hausman, 2005, p. 43–44, Hausman's emphasis, footnote ommitted)

Hausman thus agrees with Binmore that preferences should be taken as representations of what rational people do all things considered and that they are immediately

action-guiding. But he agrees with Sen that the origins of these preferences are often complex.[2] The major advantage of Hausman's unified notion of preferences is that it is compatible with the standard concept of preferences in economics and game theory. The disadvantage is that all-things-considered preferences reduce the predictive power of game theory to zero, as Hausman recognises himself: "The task of figuring out how individuals think about their strategic interactions and how they decide how to rank comprehensive outcomes [...] is ruled out of game theory. The task resides instead in a state of limbo." (Hausman, 2005, p. 47). This means we can only infer from the observation of choices which game individuals are playing. We can never take the choices as *explanandum* and the game as *explanans*. For instance, if rational agents cooperate in what appears to be a PD, we must conclude that they are actually not playing a PD, i. e. that their all-things-considered preferences induce a different game. It is therefore important to distinguish between a game form that appears to be a PD in terms of payoffs for the outcomes, and a game, consisting of a game form and preferences, that constitute a PD.

This may sound worrying, but for game theorists it is not much of a problem. Game theory is nothing but a mathematical theory of interdependent strategic choice. Game theory, correctly understood, avoids cognitive assumptions on preference formation. In this thesis, however, I have a vested interest in cognitive processes. This requires me to express myself carefully when using the term "preferences". I adopt Hausman's proposal to take preferences as all-things-considered, action-guiding preferences. However, when I define games I will work in two steps: Firstly, I start with the material (often monetary) payoffs for the outcomes in the game form. Secondly, I state the all-things-considered, action-guiding preferences agents have. Game form and preferences together define the game. For instance, I might say (as I did above), that table 1.1 states a game form with payoffs for the row and column player *in the structure of* a PD. This leads to a prisoner's dilemma game *only if* Row and Column maximise payoff. In that case, their all-things-considered preferences over the outcomes in the game form induce a PD game. Often, however, payoffs in the form of a PD as in table 1.1 do not lead to a prisoner's dilemma, because agents do not maximise payoff and their all things-considered preferences induce a different game. Arguably, these are the more interesting cases.

[2]Hausman also rejects the revealed preferences approach, in contrast to Binmore.

2.2. Social Norms

The term "social norms" is used differently throughout the literature. Cristina Bicchieri (2006) proposes a sophisticated definition. I quote it at length:

> "*Conditions for a Social Norm to Exist*
>
> Let R be a *behavioral rule* for situations of type S, where S can be represented as a mixed-motive game. We say that R is a social norm in a population P if there exists a sufficiently large subset $P_{cf} \subseteq P$ such that, for each individual $i \in P_{cf}$:
>
> **1** *Contingency*: i knows that a rule R exists and applies to situations of type S;
>
> **2** *Conditional preference*: i prefers to conform to R in situations of type S on the condition that:
>
> **(a)** *Empirical expectations*: i believes that a sufficiently large subset of P conforms to R in situations of type S;
> and either
>
> **(b)** *Normative expectations*: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S;
> or
>
> **(b')** *Normative expectations with sanctions*: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S, prefers i to conform, and may sanction behavior.
>
> A social norm R is *followed* by population P if there exists a sufficiently large subset $P_f \subseteq P_{cf}$ such that, for each individual $i \in P_f$, conditions 2(a) and either 2(b) or 2(b') are met for i and, as a result, i prefers to conform to R in situations of type S." (Bicchieri, 2006, p. 11)[3]

In this review I do not attempt a full explanation of what Bicchieri calls a "rational reconstruction" (p. 10, italics omitted) of the concept of social norms. Instead I merely want to point out some interesting core features of her complex view. Importantly, Bicchieri distinguishes between the *existence* of a social norm and the *following* of a social norm. A norm can exist without being followed because condition 2 demands only a conditional preference to follow the norm. Thus it is possible that a sufficiently

[3] Italics by Bicchieri, but numbers added by me. The explanations following in Bicchieri (2006) make it clear that Bicchieri intended these claims to be numbered in this way.

large subset P_{cf} would follow the rule R in situations S, but only under the condition that 2(a) and either 2(b) or 2(b') are met. If these conditions are not met, the social rule exists latently insofar as those agents still have the conditional preference to follow R as soon as the conditions are met.

Social norms apply to certain situations S. However, not all situations call for social norms. The function of social norms is to solve mixed-motive games, i. e. games where the interaction between players is neither a zero-sum conflict nor a perfect partnership. Examples are the prisoner's dilemma, the stag hunt game (also called assurance game) and the ultimatum game. The situational component S is important because it allows for situations where social norms apply only to some agents (the description of the situation can single out a subset of the population). It also opens up a second route how social norms can be latent: It is possible that a social norm exists, but that all agents avoid situations S in which the norm applies.

Social norms "solve" mixed-motive games by transforming them into coordination games. The coordination game has two pure-strategy equilibria: All agents comply, or all agents do not comply. An agent faces two questions here: The first question is whether the other players are conditional compliers (whether they play a coordination game). To decide this, the agent needs to form beliefs about the beliefs of other agents regarding conditions 1 and 2. The second question is whether the agent prefers to comply with the norm (given that he plays a coordination game), i. e. whether conditions 2(a), and either 2(b) or 2(b') are met. If a norm exists, all or most agents are conditional norm followers. But conditional compliance is not always actual compliance: What remains is the need to coordinate on the compliance equilibrium. Bicchieri describes the first question as a Bayesian game. In its simplest version, a norm regulates interactions between just two persons. In each interaction, an agent interacts either with a conditional complier or a violator. The problem for the agent is that he does not know whether the opponent is the former or the latter. Agents need to estimate the probability whether they are playing against a conditional complier or a defector. This is why the sufficiency requirements in (2a), (2b), and (2b') are subjective measures: Each agent has different thresholds as to what is sufficient to change his belief and decide whether he assumes to be in a coordination game or a mixed-motive game (Bicchieri, 2006, p. 223). A "suspicious" agent will assume to play against a violator unless he has a very high degree of belief that he plays against a conditional complier. A "trusting" agent has a less demanding threshold to assume that the other agent complies conditionally.

Suppose that an agent's levels of subjective sufficiency have been reached, so that her beliefs about other agents' beliefs induce her to think she is involved in a coordination game. How does she decide between "compliance" and "violation" in this coordination game? Bicchieri thinks that several motives can lead to compliance, among them self-regarding payoff-maximisation (fear of punishment, for instance), benevolence, or altruism. Thus, agents are not simple payoff maximisers. Agents form beliefs about what the other agents are going to do, and make a decision according to their personal utility function, taking these beliefs into account. Again, what is sufficient to induce compliance may vary from agent to agent. Therefore, Bicchieri's use of the term "sufficient" points to subjective thresholds that need to be reached for an agent to comply. Agents have beliefs about the number of conditional compliers, beliefs about the number of actual compliers, and beliefs about the normative expectations of others and their propensity to sanction norm violations. The necessary thresholds to switch from violation to compliance (and vice versa) differs from agent to agent and they can only be determined empirically, according to Bicchieri (p. 12).

Social norms exist within populations. There may be populations in which the norm exists, and others in which it does not. Also, for a norm to exist, it is not always necessary that all agents conform, or that all agents are conditional conformers. Since $P_{cf} \subseteq P$, it is possible that some agents do not conform, not even conditionally. This accounts for cases of ignorance, weak will, opportunism, or other motives for non-compliance. In addition, the set of agents actually following the norm can be smaller than the set of conditional followers. This can occur because agents might make mistakes, or have different thresholds of sufficiency.

Bicchieri argues that social norms differ from moral norms regarding empirical expectations: Social norms, according to Bicchieri, only exist when people have empirical expectations that many other people follow them.[4] The existence of moral norms, by contrast, does not depend on any empirical expectations regarding the level of compliance: "by their very nature, moral norms demand (at least in principle) an unconditional commitment" (Bicchieri, 2006, p. 20). She then explains that "expectations of other people's conformity to a moral rule are not a good *reason* to obey it" (p. 21, Bicchieri's emphasis). I concur with Bicchieri that if we ask for the reasons why we ought to follow a moral norm, the fact that many people obey it is not an acceptable argument. But I disagree with Bicchieri's distinction regarding the justification of social and moral

[4] In my opinion, Bicchieri makes an unnecessarily strong claim when talking about the "existence" of norms. Some norms may be latent, inactive norms in the sense that contingency is given, but not empirical or normative expectations. A latent norm could still be relevant (and "exist" in this sense) because it may be activated when expectations change.

norms. We can neither justify social nor moral norms by referring only to the number of people obeying it.[5] To justify a social norm, we must argue that obeying it is the right thing to do, because it solves the dilemma posed by a mixed-motive game. For instance, the reason why we should obey a social norm which transforms a prisoner's dilemma into a cooperation game is that we ought to obey norms that benefit the group as a whole, rather than maximising individual payoff.

The upshot is that our empirical expectations regarding norm compliance alone neither justify moral nor social norms. However, there are other norms where empirical expectations are indeed decisive. They are usually called conventions (Lewis, 1969). A convention is a rule to solve a coordination problem. Table 2.1 gives the standard example. Row and Column are drivers on a street, approaching each other. There are

		Column	
		drive left	drive right
Row	drive left	1, 1	0, 0
	drive right	0, 0	1, 1

Table 2.1.: A symmetric coordination game. The numbers are ordinal preferences.

two strict pure strategy Nash equilibria. Either they both drive on the left, or they both drive on the right.[6] If they cannot coordinate on one of these, the outcome is disastrous for both.[7]

[5] Interestingly, Bicchieri signals readiness to concede that our moral norms are conditional (p. 20, n. 10). In a Hobbesian state of nature, even the rule not to kill may not hold, given that self-defense is necessary. However, one should distinguish whether a norm can be justified, or whether a norm violation can be excused. I prefer to say that killing can never be justified, but it can be excused in certain extreme circumstances.

[6] In addition, there is a mixed strategy Nash equilibrium when Row and Column randomise their choice with probablity 0.5. This pareto-inferior equilibrium is seldom stable in practice.

[7] A variation of this game is the "Battle of the sexes" (not displayed). There are still two strict pure strategy Nash equilibria, but the two players have different preferences over the two Nash equilibria. Again, there is an additional mixed strategy Nash equilibrium. The probabilities for the randomising device depend on the exact payoffs. The standard example is a couple planning to spend an evening together (and they prefer to do *something* together over doing things separately). Row wants to go to the Opera, Column to the football match. The problem is now more complicated, because the two pure strategy equilibria are of different value for the two players.

For conventions, the empirical expectation of how other people behave is crucial, not only for the existence of a convention, but also for its justification. In the driving game, it is a perfectly legitimate justification to say "I drive on the left because everyone else in this country drives on the left". The nature of a convention is to coordinate. Therefore, what is the right thing to do depends only on what most other people do.[8]

One can either ask whether a norm is obeyed in a society, or whether it can be justified. To answer the first question, Bicchieri's definition does equally well for social norms and for moral norms. To answer the second question, one should (at least initially) abstract from empirical considerations as far as possible, and should ask whether the norm is in some sense a just or good one. What makes it just or good can be answered in very different ways of course. It could the the maxim which we think should be a universal law. Or it could be the utility maximising norm. It could be the norm we choose behind a veil of ignorance, etc. No matter which moral philosophy you subscribe to, the justification of a norm is surely not dependent on the rate of people complying with it.[9]

Setting aside the issue of moral norms, I find myself largely in agreement with Bicchieri's taxonomy of norms. Nonetheless, I need a different perspective to define the norms I am interested in. Rather than looking at conditions of existence and justification, I am primarily interested in the *function* of a norm. Agreeing with Bicchieri, I think that a social norm is used to overcome dilemmas of mixed-motive games. To put it slightly differently: I am interested in norms which have the function to induce cooperation to produce collectively beneficial results. This leads me to the question of cooperation in social dilemmas.

2.3. The Problem of Cooperation

The game-theoretical analysis of repeated games differs from the simple strategic games I have described above. We now have to account for the fact that the game has more than one stage. This requires a more elaborate definition of the game. A thorough treatment can be found in every advanced game theory textbook (see for example

[8] One could argue that Hi-Lo games are an exception: If one of the pure strategy Nash equilibria is strictly pareto-inferior for both parties, one could argue that both parties should act such that they reach the Hi equilibrium. Nevertheless, it would be foolish for a single individual to choose Hi when all other people choose Lo. Not only would it hurt the individual, it also hurts the opponent who expected the individual to play Lo. Consequently, one has good reasons to do what most people do, even in the Hi-Lo game.

[9] For instance, Bicchieri thinks that norms of fairness are social norms. Would it not be weird to claim that a fairness norm is just because most people comply with it?

Osborne and Rubinstein, 1994), and I restrict myself to basic terminology for repeated games with perfect information. A repeated game arises when the same players play a strategic game several times. I call each of these strategic games a "stage game" and a strategy for one stage an action. The repeated game is now constituted by the set of players, the set of all possible sequences of actions (each sequence is one possible outcome of the game), and preferences for each player over all those sequences.[10] A strategy for a player i in a repeated game determines an action for i in each stage game that can occur for each possible histories of previous games. More informally, a strategy is a "masterplan" in which the player determines in advance what he is going to do for all logically possible ways the game can go.

The question of cooperation becomes more interesting when agents play infinitely repeated prisoner's dilemmas. For political science and related disciplines, Axelrod's "Evolution of Cooperation" is the seminal contribution (Axelrod, 1984) regarding repeated games. However, the insight that repeated games can lead to cooperation is older than Axelrod's book. This is a corollary of the so-called folk theorem (cf. Gintis, 2000*a*, pp. 126–129; Osborne and Rubinstein, 1994, pp. 149–149). A formal statement of the folk theorem is beyond the scope of this thesis, but the intuitive idea for an infinitely repeated, symmetrical two person game is simple: Consider the maximum punishment a player has against the other player: The player can choose the action that gives the worst result for the opponent (assuming that the opponent plays best response) in all following rounds. In the case of the PD, the maximum punishment is to defect in all rounds to come. The folk theorem says: Any outcome that is preferred by both player to all outcomes where the player's opponent uses the maximum punishment, is a Nash equilibrium in the infinite game. Why? Because the players can agree on a sequence of actions that both of them prefer over suffering the maximum punishment. They also threaten each other with the maximum punishment just in case the opponent deviates from the agreed sequence. With these threats in place, neither has an incentive to deviate. This "trigger strategy" of punishment demonstrates that there are infinitely many Nash equilibria in infinite games.

Axelrod's specific strategy TIT-FOR-TAT is not quite as as special as initially thought.[11] Firstly, there are strategies that systematically outperform TIT-FOR-TAT (Nowak and Sigmund, 1993). Secondly, TIT-FOR-TAT is very sensitive to trembling

[10] For infinitely repeated games, things become more complicated because the sequence of possible actions is infinite.

[11] Axelrod himself anticipated this critique. He does not claim that TIT-FOR-TAT is the objectively best strategy. He merely claims that it has certain characteristics that cause its success against many strategies.

and mistakes: If one agent accidentally defects, the other will defect in the next round, which then causes the first agent to defect subsequently, etc. Therefore, one single mistake can lead to the breakdown of cooperation with TIT-FOR-TAT (Fudenberg and Maskin, 1990). However, this does not devalue Axelrod's core insights: Firstly, in infinitely repeated non-cooperative 2-person games, cooperation can emerge. Secondly, a successful strategy should be cooperative in a conditional way, that is "reward" cooperation and "punish" defection. Thirdly, the success of strategies must be evaluated conditional on the ecology of strategies: How well a strategy does depends on its opponent strategies.

Axelrod's book started a great theoretical enterprise to explain cooperation. Biologists wondered about cooperating organisms and how animal cooperation would fit into the theory of evolution. Social scientists were interested in collective action problems and the provision of public goods (Hardin, 1982). Economists wondered whether new ideas regarding cooperation could help to bridge the gap between the *homo economicus* assumption and the empirically observed levels of cooperation in reality (Camerer and Fehr, 2006). Mathematicians and computer scientists cared about the theoretical aspects of game theory and evolutionary dynamics. The focus turned towards more realistic settings. In particular, researchers started to focus on n-person dilemmas, used evolutionary dynamics, and introduced spatial structure.

2.4. An Evolutionary Perspective on Cooperation

Nowak (2006*b*, see also 2006*a* for an in-depth treatment) describes five mechanisms for the emergence of cooperation:

1. Kin Selection
2. Direct Reciprocity
3. Indirect Reciprocity
4. Network Reciprocity
5. Group Selection

Kin selection and group selection are concepts rooted mainly in the biological literature on evolution. While the existence of kin selection is uncontroversial, group selection is highly contested. To explain both concepts, it is necessary to think about the unit

of selection in evolution. The most basic unit of selection is usually assumed to be the gene (Dawkins, 1989) because genes are what is replicated. When mutation and recombination occur, it affects the genes. Genes are stable enough to be the unit of selection, yet unstable enough to change and let evolution happen.

Kin selection explains why there is an incentive for two related animals to cooperate. Genes have a—metaphorical—interest in replication. If two organisms are genetically related, it can make sense for one to help the other. The act of helping is costly and reduces the chance of replication for the helper, but it produces a benefit for the recipient. Overall, the replicative success of the gene can increase, because there is a positive probability that the two related organisms share the same gene. Kin selection has strong explanatory power for the social behaviour of animals. For instance, kin selection can (to a certain extent) explain high levels of cooperation between insects that are highly related due to their haplo-diploid sex determination system: In an ant colony, the sterile working ants have a 3/4 relation coefficient with their sisters.[12] It is less clear to what extent humans are influenced by kin selection. While it probably plays some role, it can certainly not explain the frequency of cooperation between non-relatives.

Group selection, in contrast to kin selection, is a very controversial concept (see for instance Wilson and Sober, 1994; Dawkins, 1994). In a nutshell, proponents of group selection, or "multilevel selection", want to show that there are other units of selection, apart from the gene. This is usally motivated by the desire to explain cooperation between non-related animals. If selection also took place on the group level, then more cooperative groups could have a selective advantage. Opponents of group selection (for example Dawkins and John Maynard Smith) remain unconvinced. The core problem for the theory of group selection is to show how groups emerged as units of selection in the first place, and how they ensure that "selfish genes" in the group do not unravel co-operation. However, since this thesis is not about biological foundations of cooperation, I will not engage in these debates any further.

Direct, indirect, and network reciprocity are of greater interest for this project. We have already seen the effects of direct reciprocity when looking at Axelrod's computer tournaments and TIT-FOR-TAT. In the biological literature, direct reciprocity goes back to Trivers (1971). As mentioned, direct reciproxity is restricted to settings where the same individuals interact over a longer (potentially infinite) period of time. Indirect reciprocity, by contrast, does not need repeated interactions between the same individuals (for a review see Nowak and Sigmund, 2005 and Nowak and Sigmund, 1998

[12] Under ideal condition, with only one queen in the colony.

for a more technical treatment). Rather, it suffices to have a public track record of how agents have behaved in previous interactions. Agents are only willing to cooperate with those who have a good track record, and maybe even punish those who have a bad one. Humans follow this system when they make their cooperation conditional on a person's reputation. The internet auction platform "eBay" is the classic example: After each transaction, buyers and sellers leave feedback about each other. Only buyers and sellers with good reputation, i. e. good feedback from previous rounds, will find transaction partners in the future. This provides an effective mechanism to enforce cooperative behaviour. Each single interaction is still a prisoner's dilemma. Defection is the dominant strategy in a single round. With reputation, however, there is a "shadow of the future". A lower reputation diminishes future incomes, so that the prisoner's dilemma turns into a cooperation game.

Mathematicians and evolutionary biologists have found that indirect reciprocity produces a lot of puzzles: "The calculations of indirect reciprocity are complicated and only a tiny fraction of this universe has been uncovered", as Nowak (2006*b*) remarks. For one, there are several sensible conceptions of indirect reciprocity because there are several concepts of reputation. It is clear that reputation should go up when a "nice" agent cooperates with another "nice" one. Also, reputation should go down when a "bad" agent defects defects against a "nice" one. But what happens to reputation when both agents defect? Or when one agent defects as a punishment because the other defected in the previous round? Or, conversely, when the agent cooperates, even though the other should be punished for defecting previously (see Leimar and Hammerstein, 2001 and Ohtsuki and Iwasa, 2006 for a thorough formal treatment)?

In addition, the evolutionary dynamics of indirect reciprocity are rather complex. Again, the problems can be merely touched upon here. One issue is that a homogenous population of agents, who follow a norm of indirect reciprocity, can be invaded by agents who cooperate with everyone, i. e. agents who lack discriminating behaviour and do not defect against agents with low reputation. If the number of discriminators falls below a certain level, this population can in turn be invaded by defectors and indirect reciprocity breaks down (Nowak and Sigmund, 1998; one possible solution concept among many is Brandt and Sigmund, 2004).

Until recently, researchers have paid little attention to the influence of spatial structure on the emergence of cooperation. Notable exceptions are Nowak, Bonhoeffer and May (1994), Skyrms and Pemantle (2000) and Alexander (2003*b*, 2007), pointing out that spatial arrangements can have an important impact. Nevertheless, most research has focused either on homogeneous populations with no spatial restrictions for the in-

teractions of agents, or on lattices with interactions restrained to neighbours on a grid. Progress in the literature on networks and the application of graph theory in the natural and social sciences (see Strogatz, 2001) has now sparked a first wave of papers that explicitly consider the effect of different network structures on agent-based models of cooperation (Lieberman, Hauert and Nowak, 2005; Ohtsuki et al., 2006; Ohtsuki and Nowak, 2006*a*; Santos and Pacheco, 2006; Ohtsuki and Nowak, 2006*b*, and further references in these papers).

In this first wave of literature the structure of the network is static. The structure influences the behavior and payoffs of the agents on the network, but the agents are not able to change the social structure determined by the network. Some of the most recent papers relax this restriction and have begun to explore dynamic networks (Pacheco, Traulsen and Nowak, 2006*b*,*a*; Tomassini, Luthi and Giacobini, 2006; Fu et al., 2007). Agents can influence the agents they have contact with and thereby shape their neighbourhood. This mirrors the nature of social structures in reality: Agents have some, though not complete, control over the set of people they interact with. They can cut ties with those who cheat them and establish ties with those who seem trustworthy. Such networks can be of a professional nature (trade networks, academic collaboration, etc.) or a private one (networks of acquaintance, social networks in virtual worlds, etc.). The paper most relevant for my own research is Santos, Pacheco and Lenaerts (2006), because the authors deal explicitly with the effects of a changing interaction network. In contrast to other papers, my model in chapter 7 deals with n-person prisoner's dilemmas. To my knowledge this is the first implementation of n-person games on dynamic networks. In contrast to the papers cited earlier, I do not analyse the effects of networks in an evolutionary setting. The most important difference, however, is that this thesis attempts a philosophically informed perspective on the simulation of human social dynamics, rather than biological systems.

Apart from the increasingly technical debate regarding indirect reciprocity, the question remains whether the standard forms of reciprocity as discussed above can explain human cooperation and human altruism. Some authors argue that a different concept is needed to explain the extraordinary levels of human cooperation: the concept of "strong reciprocity" (Gintis, 2000*b*; Bowles and Gintis, 2003; Fehr and Henrich, 2003). Fehr and Henrich offer a succint definition: "The essential feature of strong reciprocity is a willingness to sacrifice resources in both rewarding fair behavior and punishing unfair behavior, *even if this is costly and provides neither present nor future economic rewards for the reciprocator*" (p. 57, their emphasis). Strong reciprocity is unlike direct or indirect reciprocity because these forms of "weak reciprocity" are motivated by

future benefits. Strong reciprocity, by contrast, motivates cooperative and altruistic behaviour even when no future interaction is going to happen and no future benefits can be gained. The term "strong reciprocity" might be misleading here because there is no further reciprocation necessary for an act motivated by strong reciprocity. Note, however, that strong reciprocity is not unconditional cooperation: Cooperation is still conditional on earlier cooperation by the other player(s).

Research into strong reciprocity is mainly motivated by empirical studies demonstrating that weaker forms of reciprocity fail to account for human cooperation in many settings. I give an overview of these results in the next section.

2.5. Evidence of Cooperation

Evidence for human cooperative behaviour has been provided by social psychologists and behavioral economists. Good starting points are the reviews by Ledyard (1994) for public goods problems, and Sally (1995) for prisoner's dilemmas. The question is no longer whether humans cooperate more often than would be expected from a *homo economicus* (the answer to this question is a firm "yes"). Rather, the challenge is to describe and understand the conditions that lead to a deviation from *homo economicus* behaviour, and to explain these findings within an evolutionary framework (see Camerer and Fehr, 2006).

The Swiss economists Ernst Fehr, Simon Gächter and Urs Fischbacher produce some of the most exciting work regarding human cooperation (for reviews see Fehr and Fischbacher, 2004; Gächter, 2006). I want to emphasise three conclusions:

1. The desire to obtain social approval can increase the level of cooperation (Gächter and Fehr, 1999; Rege and Telle, 2004).

2. A positive feeling of group identity can increase the level of cooperation (Gächter and Thöni, 2005; Goette, Huffman and Meier, 2006; Bernhard, Fehr and Fischbacher, 2006; Dawes, van de Kragt and Orbell, 1988).

3. Many people show a disposition to be "conditional cooperators", that is they cooperate as long as they believe that others cooperate too (Gächter, 2006; Fischbacher and Gächter, 2006).

These points link with my discussion of group formation and exclusion in chapters 6 and 7.

2.6. Sociobiology and Evolutionary Game Theory

Although this thesis is not primarily concerned with sociobiology or evolutionary game theory, both approaches have an impact on the field I am interested in. Evolutionary thinking influences the analysis of human psychology and norms. Some researchers have the ambition to replace traditional moral philosophy with a science of morality based on evolutionary principles. The economist Ken Binmore (1994, 1998, 2005) is a strong exponent of this position. Binmore rejects the whole tradition of normative reasoning. Instead, he devises a naturalistic research programme that is supposed to explain the emergence of justice and norms as a natural process of biological and cultural evolution.

Binmore's approach was fiercely criticised by Sugden (2001*b*), and sparked an angry reply by Binmore (2001). Put briefly, Sugden argues that Binmore fails to naturalise notions like "rationality" or "fairness". He also attacks Binmore's lack of empirical support for his theory, and his high-handed critique of *a priori* reasoning, without acknowledging that his own argument relies heavily on *a priori* reasoning.

Binmore is not the only researcher trying to "debunk" conventional moral philosophy by using evolutionary arguments (see Lillehammer, 2003). Another "debunker" is Michael Ruse (1995), who, like Binmore, argues that morality is just a useful fiction that has provided humans with adaptive fitness. Railton (2000) challenges the "debunkers" with this task:

> "Any debunker who proclaims that morality could not emerge if natural selection did not *itself* do the work of implanting within us a sense of justice, fairness, impartiality, etc. is welcome to explain how epistemology could have emerged since natural selection did not (I presume) *itself* do the work of implanting us with a devotion to truth for truth's own sake or a commitment to impartial norms of epistemic assessment." (Railton, 2000, p. 59)

If "debunking" means that all human behaviour must be reduced to a naturalistic explanation, then the "debunkers" must explain why people should care for the truths of theories at all. Naturally, "debunkers" have tried to meet this challenge. One example is Gintis's (2007) attempt to link rationality and an interest in truth to adaptive fitness.

While Binmore sees his theory as the spearhead in an all-out attack against moral and political philosophy, other authors tread much more carefully. Brian Skyrms (1996; 2004) works with iterated two-person games and replicator dynamics. Skyrms tries to understand elementary processes of evolution that might have helped to produce norms

of cooperation and equality. He espouses a conciliatory view regarding the relation between moral philosophy and evolutionary research:

> "Ethics is a study of possibilities of how one might live. Political philosophy is the study of how societies might be organized. If possibility is construed generously we have utopian theory. Those who would deal with 'men as they are' need to work with a more restrictive sense of possibility. Concern with the interactive dynamics of biological evolution, cultural evolution, and learning provides some interesting contraints." (Skyrms, 1996, p. 109)

My own position is closer to Skyrms than to Binmore and Ruse. These issues will be taken up in chapters 3 and 5.

2.7. Social Dynamics and Computer Simulations

The literature on computer simulations in the social sciences is growing rapidly (see Gilbert and Troitzsch, 2005; Tesfatsion, 2007, with further references). In this short review I focus on agent-based models. A general overview is provided by Page (2005). Miller and Page (2007) discuss foundational issues and give examples. Agent-based models can help social scientists to extend their analysis to phenomena that are difficult or impossible to observe with other methods. Their strength lies in modelling bounded rationality (Selten, 2001), the dynamics of social interaction (rather than just the outcome), the complexity resulting from agent interaction, and the effects of social space.

There are by now some impressive applications of agent-based modelling in economics, political science, and theoretical biology. What is largely lacking, however, is a philosophy-of-science framework to explain the methodological status of agent-based models. One notable exception is Scott de Marchi (2005), who discusses the merits of computational models in comparison with analytical techniques, particularly game theory. De Marchi identifies the "curse of dimensionality" as the main challenge of model building in the social sciences. A model rests on dozens of explicit and implicit assumptions, and it is often questionable whether a slightly different model would have produced the same results. If models are underdetermined by data, the use of these models may be dubious. Computational modelling, according to de Marchi, has the advantage that scientists can explore the parameter space of the model systematically.

With a computer program, tests like "what would happen if I assume that..." are easy to carry out.

De Marchi (2005) discusses the use (and limitation) of computational models when deductive game-theoretical reasoning fails. Discussing Deep Blue's success against chess grandmaster Garry Kasparov, de Marchi notes that Deep Blue did not "solve" the game of chess with backward induction, and, indeed, computational complexity rules this out. Rather, Deep Blue calculated future positions in a limited number of moves and then used a function to evaluate these positions. Even with brute computational force it is impossible to calculate a game of chess to its end. While the utility function of chess has only three elements ({win, draw, lose}), Deep Blue has to use "idiosyncratic utility functions ... comprised of features that bear no necessary relation to the utility function or rules of chess but are nonetheless helpful in evaluating intermediate positions in chess" (de Marchi, 2005, p. 85). In other words, Deep Blue splits the game of chess (and its theoretically existing complete extensive form) into manageable components and plays the game by evaluating components quite independently from its complete extensive form.[13]

The example is instructive because it shows that a working computational strategy for solving complex games is quite similar to the way humans approach complex problems. Chess grandmasters use highly sophisticated position evaluation procedures, together with a limited capacity to compute future moves (offset by an excellent ability to prune irrelevant sequences of moves in their computations). The same strategies are applied in other complex games (Go, Bridge, etc.), and most likely in many complex social interactions. When human agents make decisions they weigh their options by using evaluation heuristics rather than going through the whole extended form of the game at hand. If the game is too difficult to be solved analytically, real agents will use heuristics, and so should the agents in computational models (Gigerenzer, 2001).

In practice computational models suffer from the curse of dimensionality, too: The number of assumptions in a computational model is enormous. Think about the number of decisions made when programming even a relatively simple model. Even worse, there is a strong temptation to adapt the model to obtain the results one wants to see, especially since it is difficult for outsiders to gain a full understanding of how the program works in detail. This is particularly true for large simulations with many assumptions. For this reason, I am sceptical regarding large-scale simulations (see

[13]The recent draw between Deep Fritz and grandmaster Vladimir Kramnik fits into de Marchi's explanation. Deep Fritz had only 1.3% of the computational power available to Deep Blue, but a much more sophisticated evaluation function (Kurzweil, 2006).

Kliemt, 1996), as discussed in chapter 4.

There are fundamental differences in philosophy of science regarding agent-based models. Epstein and his collaborators argue for a "generative social science", trying to "grow" social phenomena from the "bottom up" (Epstein, 2006). The idea is to explain social phenomena on the macro-level by creating dynamic models based on assumptions about the micro-level. Grüne-Yanoff (2006) points out that "the *generandum* is the *explanandum*" (p. 5) in this approach. The model is supposed to explain because it is able to reproduce reality. If the model behaves like a real society, then—it is claimed—we know the causes of this behaviour (because we have built all causes into the model). This view is mistaken, as Grüne-Yanoff argues. Firstly, computational models, even the most sophisticated, are still radical simplifications. Therefore, they only match reality with regard to specific variables. Secondly, models with deliberately unrealistic assumptions often have better predictive power. One would hardly argue that such models are the *explanans* of reality. Thirdly, there is not one, but a whole class of models that would reproduce the same time series of relevant variables. Which of these models should we choose?

Grüne-Yanoff argues that the *explanandum* of computational models should not be the *generandum*. Rather the *explanandum* is the disposition of the system modelled. This entails giving up causal explanation and embracing constitutive explanation: "*Causal explanation* explains an event by citing its predominant cause. *Constitutional explanation* explains a system's disposition by showing how the system's elements have properties that constitute it." (p. 17, Grüne-Yanoff's emphasis). This links Grüne-Yanoff's argument with Sugden's article on the explanatory merit of theoretical economic models (Sugden, 2000). Sugden discusses two classic theoretical models (Akerlof's "Market for Lemons" (1970) and Schelling's "Segregation Model" (1978)). He finds that both models make counterfactual assumptions. He rejects an instrumentalist interpretation, as the models do not make any directly testable predictions. Instead, Sugden argues, these models should be understood as "caricatures" describing "tendencies" of reality. These tendencies are meant to be broad generalisations: they hold over a wide range of *ceteris paribus* assumptions. How can simple models make such bold claims? Sugden shows that these models require an inductive leap from a very specific causal relation in the model to a wide range of causal relations in the real world. And how can this induction be justified? Sugden offers a two-part answer: "robustness" and "credible worlds". Regarding the first part, Sugden writes: "Robustness arguments work by giving reasons for believing that a result that has been derived in one specific model would also be derived from a wide class of models, or from some very general

model which included the original model as a special case." (p. 22). Thus, a good model is not merely a single model, but contains a whole class of models. Regarding the second part, Sugden argues that a model tells us something about the real world if it is an instantiation of how the world could be. The model is a caricature, but a good caricature could be the truth in a not too distant possible world. I discuss and criticise Sugden's view in greater detail in chapter 4.

Sudgen's article refers to the more general question of representation in the philosophy of science. Frigg (2006) puts the question well:

> "Models are representations of a selected part or aspect of the world [...]. But in virtue of what is a model a representation of something else?" (Frigg, 2006, p. 50)

Recent attempts to answer this question deny that there is a relation of isomorphism between model and world and look for alternative conceptions (Giere, 1988; Weisberg, 2003; Godfrey-Smith, 2006). The issue of representation is also closely linked with the problem of idealisation in model building (Strevens, 2007; Jones, 2005; Teller, 2001; Mäki, 1992). Most successful models simplify reality by omitting or distorting certain facts. They make counterfactual or false assumptions to be successful as models (for a nice example see Küppers and Lenhard, 2005). In the natural sciences, idealisation is often useful to keep the models simple enough for analysis (Batterman, 2002). Similarly, the social sciences and philosophy also benefit from simple models to explore social mechanisms, especially when the underlying social processes are not well understood (Kliemt, 1996). This issue will be addressed in detail in chapter 4.

Part I.

Social Norm Enforcement, Modelling, and Evolution

3. Subtle Ways of Norm Enforcement: Towards a Positive Theory of Moral Constraints

3.1. Realistic Moral Theory

Moral philosophy usually focuses on desirability, not on feasibility. However, since "ought implies can", moral philosophy should not only deal with what we should do, but also with what we can do. The interpretation of the "ought implies can" principle is contested. Problems arise when agents are able to limit their options in advance (Sinnott-Armstrong, 1984): Say that I have borrowed money from you and promised to pay you back today. Before I return the money to you, I spend all the money I have in an expensive restaurant. Now I cannot return your money. Does it follow that I ought not return your money? Surely not. Sinnott-Armstrong argues that "ought" neither presupposes nor entails "can". Rather, the principle only expresses a conversational implication.[1]

For my purposes, a stripped down, more defendable version of the "ought implies can" principle will do: The fact that we are unable to perform an act *supports* the claim that we do not have an obligation to perform this act. "Support" means that the implication of the "ought implies can" principle is conditional on several restrictions. For example, "ought" does not imply "can" when the relevant agent is responsible for his inability to perform the act required. The point of this chapter is not to discuss all conditions necessary to make sense of the "ought implies can" principle. Rather, I want to work with a minimal definition. The minimal definition acknowledges that the inability to perform an act is not always sufficient to void an obligation. However, at the very least, if someone is unable to act, this is *prima facie* evidence that he has no obligation to do so. Moreover, someone who cannot perform an act is usually not held responsible for failing to perform this act, and he is not blamed for failing.

[1] Streumer (2003) proposes a new "tensed" formulation of the "ought implies can" principle to iron out the problems Sinnott-Armstrong raised.

3. *Subtle Ways of Norm Enforcement: Towards a Positive Theory of Moral Constraints*

The "ought implies can" principle is mainly cited in the *normative* debate regarding moral responsibility. However, there is also a much less noted *descriptive* aspect to it. If we are interested in *describing* moral systems, the "ought implies can" principle provides us with information on the constraints for morality. Moral systems are factually constrained by what moral agents can do. It is pointless to prescribe actions which can never be performed. Therefore, moral systems are bound by the factual limitations for actions. And while moral systems can be of an utopian nature, it is unlikely that common sense morality prescribes more than what can be expected from the average agent under normal circumstances. Since moral systems solve problems that human agents need to solve, it would be strange if they did not take the circumstances of action into account. Therefore, when describing and analysing morality, it is interesting to look at the constraints given by the world as it is.

Surprisingly little has been said about how the "oughts" of moral theory depend on the restrictions given in the real world. Rawls famously distinguishes between ideal and non-ideal moral theory (Rawls, 1999*b*, p. 216–218).[2] Unfortunately, he is not very precise in his explanations of non-ideal theory (Phillips, 1985). He distinguishes between "the natural limitations and accidents of human life" and "historical and social contingencies" (Rawls, 1999*b*, p. 215). I am interested in the latter problem, particularly the case of non-compliance. Rawls is very optimistic regarding compliance: "But men's propensity to injustice is not a permanent aspect of community life; it is greater or less depending in large parts on social institutions, and in particular on whether these are just or unjust." (Rawls, 1999*b*, p. 215). I agree with Rawls that the willingness to comply with relevant norms can be changed for the better, and that social institutions play a crucial role. But I think that we require a much better understanding of the social circumstances and dynamics that enable agents to comply with norms.

Let me express the idea of realistic moral theory more systematically. A moral prescription demands certain forms of human behaviour. Sometimes, the prescription also demands certain beliefs and desires because for some moral theories it does not only matter what one does, but also why one does it. Inspired by Christian List's (2006) use of the possible worlds terminology to analyse positive and normative laws, I take moral prescriptions as "modal desiderata" and use possible worlds terminology to analyse the "can" in "ought implies can" more systematically. List interprets normative laws as follows:

> "Formally, a proposition represents a normative law if it states a certain

[2] Rawls speaks more specifically of nonideal theory in "The Law of Peoples" (1999*a*), but he only deals with relations between societies there.

> fact that is true in all *permissible* worlds, relative to a given standard of permissibility. There are at least two ways of looking at this definition. We can either let the law be given and then determine the class of worlds that are permissible according to that law (this class would include all those worlds in which whatever is asserted by the given law is true), or we can let the class of permissible worlds be given and then determine what laws would single out precisely those worlds as permissible (the resulting laws would assert propositions that are true in all those permissible worlds). Depending on how we demarcate the class of permissible worlds (that is, depending on our interpretation of 'permissibility'), a law can be a constitutional, ordinary legal, or moral one.
>
> [...]
>
> A common feature of all these examples of normative laws is that they can all be interpreted as modal desiderata, that is, facts that are true across all relevant permissible worlds, although different normative laws are based on different standards of permissibility. Whereas positive laws become more robust as the relevant class of possible worlds grows, normative laws become more restrictive as the relevant class of permissible worlds shrinks: the fewer worlds are deemed permissible, the more restrictive is the law." (List, 2006, p. 206)

Following List, a prescription is a division of possible worlds: The *permissible worlds* are those worlds where all agents behave according to the prescription and entertain the prescribed beliefs and desires. The *impermissible* worlds are all worlds where some or all agents violate some or all prescriptions.[3] How large the set of permissible and impermissible worlds is depends on the moral prescription: Demanding prescriptions yield small sets of permissible worlds, less demanding ones large sets.

The general strategy of my argument is as follows: I will firstly define the set of permissible worlds and the set of possible worlds. I will then introduce the notion of the actual world and the set of feasible worlds. The set of feasible worlds consists of all worlds that are possible transformations of the actual world. I will then argue that a precise definition of feasibility is crucial for these transformations. A wide definition entails a liberal sense of feasibility, leading to a more utopian moral theory, a narrower definition a more restricted one, leading to more realistic moral theories.

Following List's definition of the set of permissible worlds, I start with some set of

[3]Similar ideas where developed in the area of deontic logics. See in particular McNamara (2006).

moral prescriptions R containing one or more moral rules. Each rule must state at least one modal desideratum of the form "it ought to be the case that ψ", with ψ being a descriptive proposition. I write $\bigcirc$ for the ought operator such that "it ought to be the case that ψ" can be written as $\bigcirc\psi$. Typically, these prescriptions refer to all individuals in a general way (e. g. "It ought to be the case that no one kills anyone"), but other forms are possible. For example, the prescriptions could refer to a limited group of persons ("It ought to be the case that Kai does not eat strawberries on Sundays") or they could prescribe beliefs and desires rather than actions ("It ought to be the case that Adam loves his brother"). Thus, all prescriptions of the form $\bigcirc\psi$ are instances of moral prescriptions.[4]

The set of all possible worlds Ω is the most encompassing set. It contains all possible worlds in which the moral theory represented by R is true. This means that Ω does not contain possible worlds with moral rules that contradict R or with rules that are more or less demanding than R. Thus, Ω is the set of possible worlds with identical moral systems R. My approch differs from the standard approach in deontic logic (see for example McNamara, 2006), where moral theories are world-bound. In standard deontic logic, different possible worlds have different moral requirements, and across worlds these requirements can differ or contradict each other. For instance, one possible world in standard deontic logic is the world in which all negations of R are true, an "anti-moral" world from the perspective of the world in which R holds. While this generality is useful for deontic logic, it is not for the aims pursued in this chapter: Here I am interested in examining the feasibility of a *given* set of norms or a moral theory. For the purposes of my analysis I want to keep R fixed across all possible worlds. However, in the next section I will discuss an option to allow for world-bound moral theories.

Apart from the restriction that R and only R is the set of moral norms true in all possible worlds, I use a rather generous definition of possibility: Possible worlds are all worlds that are logically possible.[5] Possible worlds can be radically different from our world. The requirement for a world to be possible is that its description does not contain logical contradictions. For example, the propositions p and $\neg p$ cannot both be

[4] I remain agnostic about the demarcation between moral rules and other rules for the purposes of this chapter. Therefore the definition of a moral rule is as open as possible: All sentences of form "it ought to be the case that ψ" are admitted. One could easily restrict the domain of moral rules. Which restriction one prefers depends on which moral theory one subscribes to. Typical conditions for moral rules are that they make claims about interpersonal relations, that they are (in some defined sense) impartial, that they are derived through a certain deliberative process, that they serve certain functions, etc.

[5] The exact definition of possibility does not affect the argument. It is only important that Ω contains all worlds that are deemed possible in the sense that matters (perhaps conceptual, logical, metaphysical, physical, etc.).

true at the same time in any possible world. I take it that the moral prescriptions we usually deal with *can* be realised in some possible worlds.

The important point is that the set of possible worlds can be divided into those worlds where the modal desiderata are satisfied (i e. where the prescriptions are obeyed), and those where the modal desiderata are not satisfied (i. e. where the prescriptions are not obeyed). We call the former the set of permissible worlds Π, and the latter the set of impermissible world Π^C. If $\bigcirc\psi$ is a moral prescription in R, then ψ is true in all permissible worlds ω contained in Π, or formally $\bigcirc\psi \in R \Rightarrow \forall\, \omega \in \Pi : \psi$ is true at ω.

Within the set of possible worlds we find the actual world α—the world as it is right now. The actual world is contained either in the set of permissible worlds, or in the set of impermissible worlds. If the actual world is in the set of permissible worlds, we are in the happy state that everyone complies with the norms in R. Needless to say, this happy state does not often obtain if the set of norms R has any bite. What about "almost permissible worlds" where, say, all but one person complies with the norms? These are still, strictly speaking, impermissible. However, we could extend the set of permissible worlds by using a less demanding set of norms R. For instance, the rule could only demand that 95% of moral agents comply.

Finally, we need to introduce the set of feasible worlds Φ to complete the argument. To determine the set of feasible worlds, one starts from the actual world α. Speaking generally, a world is feasible if it can be reached by changing the actual world. But how do these changes take place? Two different processes matter: Firstly, the moral agents in the world can change the world through their actions. Secondly, the world changes because of other contingent events beyond the agents' control. This roughly distinguishes changes that agents are in principle able to control, and ones they aren't. For instance, individuals can control their personal fuel consumption, but they cannot control volcanic eruptions.

The important point is that some possible worlds are difficult to produce, given the state of the actual world. We need to know how different these possible worlds are from the actual world. Therefore, we need to talk of the *proximity* of worlds. I take proximity as a measure of the difficulty to transform one world to another. Worlds are close to α if it is easy for the moral agents to transform α into these worlds. Worlds are distant from α if it is difficult or impossible to transform α into these worlds.[6] For example, it is (supposedly) easy to change the actual world into one in which people

[6]The notion of proximity or similarity between possible worlds is difficult to define. Lewis (1979) attempts to offer weights for a measure. My notion of proximity refers to the ability of agents to transform worlds, while Lewis is interested in maximally similar yet counterfactual possible worlds. Therefore, Lewis's weights do not apply here.

switch off unnecesary lights. It is difficult to create a world with a neutral CO_2 balance of human activities. It is, as far as we know, impossible to create a world in which eight billion people each have their private jet. This shows that there are different notions of feasibility, depending on what we take as doable or impossible.

Feasibility is a matter of degree (Cowen, 2007). The possible worlds terminology allows us to put different concepts (i. e. standards) of feasibility into a partial order. The idea is that a concept of feasibility is wide when it implies a large set of feasible worlds Φ. It is narrow if it implies a small set of feasible worlds Φ. A set of feasible worlds Φ_1 represents a strictly wider definition of feasibility than a second set of feasible worlds Φ_2 if $\Phi_2 \subset \Phi_1$. In this case, Φ_1 does not only contain all the worlds in Φ_2, but also at least one additional world which was not deemed feasible under the concept of feasibility represented by Φ_2. Thus, I can define a

(Definition) Partial Ordering Function for Feasible Worlds. For any two sets of feasible worlds Φ_1 and Φ_2, representing different concepts of feasibility, Φ_1 represents a *strictly wider concept of feasibility* than Φ_2 if $\Phi_2 \subset \Phi_1$.

In particular, any meaningful definition of feasibility should imply that Φ is larger than the set containing only the actual world α. This expresses the idea that some changes to the actual world must be feasible. However, not all concepts of feasibility can be ordered from wide to narrow because their associated sets of possible worlds need not be proper subsets of each other.

Here are some possible concepts of feasible sets:

(1) All transformations of α that can be reached with all physically possible human behaviour and all physically possible contingent processes (feasible set Φ_1).

(2) All transformations of α that can be reached with all physically possible human behaviour, everything else equal (feasible set Φ_2).

(3) All transformations of α that can be reached with individually plausible human behaviour, everything else equal (feasible set Φ_3).

(4) All transformations of α that can be reached with socially plausible human behaviour, everything else equal (feasible set Φ_4).

Definition (1) is the widest definition of feasibility. Formally, $\Phi_2 \subset \Phi_1$, $\Phi_3 \subset \Phi_1$ and $\Phi_4 \subset \Phi_1$. In (1), all possible paths of the future that can be brought about by human action and other contingent processes are taken into account. This is not

an appropriate definition of feasibility because the notion of feasibility should capture what can be achieved by intentional action. Sometimes contingent processes work in our favour and we have *good luck*, achieving what seemed impossible. But luck does not entitle use to say that what we achieved is feasible *under normal circumstances.* Also, from a descriptive perspective, it is likely that functioning moral systems contain prescriptions which are achievable under normal conditions for the average agent, not merely as a result of luck.

Definitions (2) to (4) exclude the influence of contingent processes by keeping "everything else equal". Strictly speaking, "keeping everything else equal" is not possible. Contingent processes always take place over time; there is simply no way to stop them. So what I mean by "keeping everything else equal" is a process of interpolation, or of hypothetically controlling for the circumstances: I assume that the contingent processes in the world move on as they did in the past. The point is that the moral agents neither face extraordinarily adverse nor helpful events, which would obstruct or promote their plans.

The focus is now on the more subtle distinctions between definitions (2), (3), and (4). I think that all of them have some *prima facie* plausibility as definitions of feasibility. Definition (2) is broader than (3) and (4) (implying $\Phi_3 \subset \Phi_2$ and $\Phi_4 \subset \Phi_2$). It takes all physically possible forms of human behaviour into account, but does not consider the psychological plausibility of different actions. In a way, this definition makes sense: Since moral agents can change, it is problematic to define in advance what is psychologically feasible for them. Defenders of definition (2) can argue that any behaviour that can be performed in a physical sense is—in principle—feasible behaviour. Defining feasibility in such a liberal way implies taking behaviours as feasible even though they are psychologically implausible. We can imagine a world in which all agents behave like saints. Such a world is certainly—in a wide sense—*feasible*, given the actual world. The agents have to mend their ways and transform into perfect moral agents. But if we define feasibility in this wide sense, the distinction between utopian and realistic theory collapses. It is certainly useful for utopian theory to think about the best worlds, and we should strive to come close to them. Unfortunately, however, we have to deal with real people, not perfect moral agents. Therefore, we have to consider settings where moral agents have to interact with other agents who are morally weak, follow different morals, or do not care for moral considerations at all.

Jean-Jacques Rousseau famously wants to take "men as they are". This leads to definitions (3) and (4), where anthropological and psychological factors are taken into account. Definition (3) considers individual psychological constraints. For instance, this

notion of feasibility could incorporate that humans are "morally weak" and are unlikely to make large sacrifices in order to help others. Such a notion of feasibility would take into account that human behaviour is driven, at least partly, by self-regarding behaviour and that pure altruism is rare. Further support for definition (3) could come from evolutionary psychology, informing us of possible psychological modules "hardwired" in the human brain in the evolutionary history of *homo sapiens* and its ancestors (Tooby and Cosmides, 1992).

One could argue that psychological constraints should not matter for normative moral theory because the point is to determine what is just, not what is easy to do. My reply is twofold. First, I am mainly concerned with a descriptive theory of morality, and for a descriptive theory psychological constraints do matter. Second, even for normative theory it must surely matter how difficult it is for an agent to obey a moral prescription: We are likely to excuse failures to comply if complying is extraordinarily difficult. When it comes to moral responsibility, psychological constraints do play a role.

The distinction between definitions (3) and (4) needs to be discussed in greater detail. I will argue for definition (4) in the next section. However, before I do that I want to relate the definitions of feasibility with the possible worlds terminology. Recall my earlier claims of the relation between possible, feasible and permissible worlds:

1. The set of permissible worlds Π, derived from the set of prescriptions R, is a subset of the set of all possible worlds Ω, formally: $\Pi \subseteq \Omega$.

2. The set of feasible worlds Φ is a subset of the set of all possible worlds Ω, determined by a suitable definition of feasibility as proximity to the actual world, formally: $\Phi \subseteq \Omega$.

Figure 3.1 displays the set relations graphically. Ω is the set of all possible worlds. Π, Φ, and $\{\alpha\}$ are subsets of Ω. Π is the set of permissible worlds, defined by the set of moral prescriptions R. Φ is the set of feasible worlds, while α is the actual world. α always lies in Φ because the actual world is always feasible.

To figure out whether a moral prescription is realistic or utopian, we need to know whether at least one of the permissible worlds is feasible. If the intersection of permissible and feasible worlds $\Pi \cap \Phi$ is empty, then the moral prescription is utopian. If $\Pi \cap \Phi$ is non-empty, then the moral prescription is realistic. Using these concepts I can define utopian and realistic moral theory:

Utopian Moral Theory. A moral theory is *utopian* if the intersection of its permissible worlds with the set of feasible worlds is empty.

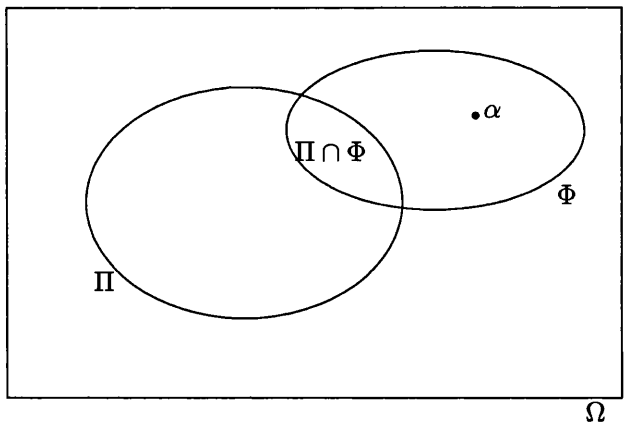

Figure 3.1.: The set of possible, permissible and feasible worlds.

Realistic Moral Theory. A moral theory is *realistic* if the intersection of its permissible worlds with the set of feasible worlds is non-empty.

How the border between utopian and realistic moral theory is staked out depends on the definition of feasibility. Wider definitions like (1) and (2) lead to a more encompassing definition of realistic moral theory. I think that (3) and in particular (4) lead to a more suitable demarcation of realistic and utopian moral theory.

It is also possible to produce a partial ordering of moral theories from realistic to utopian, condititional on a set of feasible worlds Φ. Let moral theory one be represented by the set of permissible worlds Π_1, and moral theory two by Π_2. The first moral theory is strictly more realistic than the second if and only if the intersection $\Pi_1 \cap \Phi$ contains the intersection $\Pi_2 \cap \Phi$ as a proper subset, given a specific set of feasible worlds Φ. The ranking is only partial because it is not necessary that one of the intersections is a proper subset of the other. This leads to a definition for a partial ordering function of moral theories according to their level of realism:

(Definition) Partial Ordering Function for Moral Theories according to Realism. For any two permissible worlds Π_1 and Π_2, conditional on a set of feasible worlds Φ: A moral theory represented by Π_1 is *strictly more realistic* than one represented by Π_2 if $[\Pi_2 \cap \Phi] \subset [\Pi_1 \cap \Phi]$.

A world where all agents are saints is a possible world (it is contained in set Ω), but it is not a feasible world (it is not contained in set Φ). Therefore, moral prescriptions which require a population of saints are part of utopian moral theory. The intersection

of Π and Φ is then empty. In this thesis I am concerned with realistic moral theory. The purpose of realistic moral theory is twofold. Firstly, it is a descriptive tool to analyse which norms can be obeyed under realistic circumstances and how these norms are enforced. Secondly, it is also a normative tool, because it informs us what one can realistically demand from moral agents in a social world.[7] In realistic moral theory we have to deal with situations where norm compliance is not perfect, or where different people obey different norms. Thus, norm compliers are challenged by norm violators. In these realistic settings, the set of feasible worlds (given the actual world) is smaller.

3.2. Three Extensions

The framework developed so far is very simple. In this section I propose three extensions to account for more complicated settings. The first extension allows for world-bound moral theories. The second extension indexes permissibility over agents. This extension allows us to model situations in which prescriptions are directed towards specific agents, and it enables us to differentiate between rules that require compliance by everyone and rules that require compliance by someone. The third extension introduces a distinction between norm compliance and norm acceptance.

The first extension introduces the option of world-bound moral theories. I have said above that I keep the set of moral rules R fixed across all possible worlds because I wanted to focus on the feasibility of a given moral theory. However, there is another option I have ignored until now: It is possible to keep a moral theory fixed over all possible worlds, while the prescriptions of this moral theory vary in different possible worlds. This means I assume that the moral theory, call it T, is true in all possible worlds, but the prescriptions of this moral theory are now indexed by worlds. Thus we write R_ω for all moral rules prescribed by T in ω. In our moral practice, we do not really operate with moral theories that have prescriptions ready for all possible worlds. Moral theories are usually made for the people we are and the circumstances we find ourselves in. In other words, moral theories are made (at best) for a small subset of all possible worlds.[8] However, within this subset it makes sense to say that moral theories are world-bound. With world-bound moral theories, agents have two different

[7] On the one hand, this may scale down the expectations. On the other hand, it also sets a more robust standard of what can be expected. Realistic moral theory makes it harder to excuse non-compliance by referring to unrealistic demands or difficult circumstances.

[8] How large is the subset? Different moral theorists make different claims on how universal their theory is. A Humean thinks that morality is highly dependent on the circumstances of justice, while some Kantians claim that their theory applies to all worlds with rational and reasonable beings.

options to comply with the prescriptions of the moral theory: they can either move to a world with the same prescriptions (as in the actual world) in which they comply with these prescriptions, or they can move to a world with different prescriptions in which they comply with those. Consequently, the set of permissible worlds is $\Pi = \{\omega \,|\forall \bigcirc \psi \in R_\omega : \psi \text{ is true at } \omega\}$.[9]

Oftentimes moral norms apply to specific agents or require actions from some, but not all agents. Norms can single out agents in specific roles or circumstances. For instance, there might be specific norms for the Queen of England (singling out a person by her role) or a rule for all people standing next to ponds with a drowning child inside (singling out agents by specific circumstances they find themselves in). Norms can also demand that some action is performed by a subset of agents, but not by all agents. For example, a group of nurses could be collectively responsible to ensure that their patient receives exactly one dosage of beta-blockers every day. This must be distinguished from the prescription that each nurse should give the patient beta-blockers. One way to account for personal norms is to include ascriptions to persons into the ψs. For instance, "It ought to be the case that Kai does not eat strawberries on Sundays" is personal in this way. However, a more sophisticated way to deal with these cases is to relativise prescriptions to agents. Formally, this can be achieved by indexing: The personal prescription 'It ought to be the case that agent i brings about ψ' is written as $\bigcirc\psi(i)$. One can now define the scope of i appropriately, e.g. as "Queen", "the best swimmer next to a pond with a drowning child", or "exactly one nurse of the three nurses responsible for patient P".

If we relativise prescriptions to agents, we should distinguish the sets of permissible and feasible worlds for all agents together (Π), and the sets of permissible and feasible worlds for each agent i individually. The set of permissible worlds $\Pi(i)$ for an agent i consists of all worlds in which i complies with all impersonal and all personal norms directed to him. The sets of permissible worlds for any two agents will usually differ, given that many norms demand different things from different agents. Similarly, we can introduce individualised notions of feasibility, such that all possible worlds feasible for agent i are $\Phi(i)$. To determine what is feasible collectively, we need to consider the dynamics between agents and how the different actions affect each other, as I will argue

[9]This approach can be made more general by conditioning moral theories T to the actual world α, and the set of permissible worlds to the moral theory as it obtains in the actual world, that is $\Pi_{T(\alpha)}$. This captures the idea that different moral theories (and not just different prescriptions) can be true in different possible worlds. A full development of this approach is beyond the scope of the thesis.

below.[10]

The third extension introduces a distinction between acceptance and compliance. A full treatment of this distinction is beyond the remit of this thesis. For now I merely want to point towards some interesting additions to my framework. When we look at moral norms, we often find that people accept certain rather demanding norms in principle, but perform actions according to a less demanding, but related set of norms. One could say that the accepted norms are ideals, while the less demanding, applied norms are practical principles. These practical principles are less ambitious than the ideals, but they are at least guided by or related to these ideals. We can now distinguish between two notions of feasibility: acceptance-feasibility and compliance-feasibility. A norm is acceptance-feasible if agents can accept that norm (relative to a certain standard of acceptance-feasibility). A norm is compliance-feasible if agents can comply with this norm (relative to a certain standard of compliance-feasibility). Under normal circumstances, the set of compliance-feasible worlds is a subset of the acceptance-feasible world, because it is easier to accept a norm than complying with it.

One interesting question is how strong the notion of acceptance should be. It could reach from a mere lip-service to a willingness and readiness to defend a social norm in

[10]It is interesting to think about possible aggregation functions to derive sets of collective permissibility and feasibility from the individual sets. It seems plausible to assume that collective permissibility and feasibility should be a function of individual permissibility and feasibility across all agents. The set of permissible worlds for a society as a whole could be the intersection of all individually permissible worlds, such that $\Pi = \bigcap_{i=1}^{n} \Pi(i)$ for n individuals. Consider the nurses example: Each nurse can either dispense the medication (D) or not dispense (N). Let there be three nurses. We can describe a world as an ordered tuple $\{D, N\}^3$, such that the first element stands for the action of nurse 1, the second for nurse 2, the third for nurse 3. The set of all possible worlds (with regard to medication) is $\Omega = \{NNN, DNN, NDN, NND, DDN, \ldots, DDD\}$. The set of permissible worlds for nurse 1 is $\Pi(1) = \{NDN, NND, DNN, NDD\}$. The nurse dispenses when no one else has, and does not dispense when someone has. Why does $\Pi(1)$ contain NDD? Because it is still permissible for nurse 1 not to dispense, even though NDD is collectively impermissible. Nurse 1 does not do anything wrong here, nurse 2 and 3 do. Note that this formalisation requires nurses to change their course of action depending on what the other nurses did. Thus, which action is permissible or impermissible needs to be determined dynamically. Note that the intersection of all individually permissible worlds leads to the correct set of collectively permissible worlds $\Pi = \{DNN, NDN, NND\}$.

While we can take the intersection of all individually permissible worlds to determine the set of collectively permissible worlds, it is not clear to me whether something similar works for the individual and collective sets of feasible worlds. The intuituitively plausible approach is that the union of individually feasible worlds is the set of collectively feasible worlds ("what we can do together is the sum of the things we can do on our own"). But that is not true: What we can do together is often more than the sum of our individual actions. And sometimes, what we can do together is less than the sum of what we can do individually. It is not feasible for any single person to push my broken-down car to the workshop, but it is possible for four or more persons *together*. It is feasible for you to eat this apple, and it is feasible for me to eat this apple, but it is not feasible for *both* of us to eat this apple. It seems to me that the aggregation from individually to collectively feasible sets is more complex, and requires a context-sensitive, dynamic analysis.

public deliberation. If the notion of acceptance is wide, we introduce room for hypocrisy because the gap between ambitious accepted norms and much less ambitious practical principles can grow large. A richer notion of acceptability seems preferable. Acceptance should entail a certain level of consistency within an agent's system of norms, and a willingness to argue for a norm.

Ideals are often utopian in terms of compliance-feasibility, but realistic in terms of acceptance-feasibility. Despite being utopian in the compliance sense, ideals can have an important function for the realistic principles. Imagine that without accepting the (utopian) ideals, agents would not be able to comply with the (realistic) practical principles. This looks like a faithful description of human moral psychology, as we often find a gap between ideals and actual behaviour on the one hand, but a need for ambitious ideals to control human moral practice on the other hand. In such cases, it would be wrong to dismiss the ideals as utopian (in a bad sense), because they are important to ensure the feasibility of the less demanding practical principles. The distinction between acceptance and compliance makes room for the guiding nature of ideals and avoids a cynical rejection of ideals as unrealistic. It also allows us to distinguish between three forms of feasibility:

1. Acceptance-feasibility without compliance-feasibility;
2. Acceptance-feasibility and compliance-feasibility;
3. Compliance-feasibility without acceptance-feasibility.

A good example for the first case is the acceptance of ideals, as discussed above. Acceptance is feasible insofar as agents can publicly argue for their ideals, but it turns out that an implementation is not feasible. In the second case, the norm can be fully accepted and implemented. Here we can say that the actual practice is backed up with normative acceptance. In the last case, it is feasible to comply, but it is unfeasible to accept the norm, in the sense of defending it publicly. For instance, if people are ashamed to obey a certain norm, but the costs of non-compliance are high, they might find it feasible to comply, but unfeasible to accept the norm whole-heartedly.

3.3. Social Constraints for Realistic Moral Theory

I now want to take up the adjudication between (3) and (4) as definitions of feasibility. To repeat, the two definitions are:

(3) All transformations of α that can be reached with individually plausible human behaviour, everything else equal.

(4) All transformations of α that can be reached with socially plausible human behaviour, everything else equal.

I believe that the fourth definition of feasibility is the most useful for developing a realistic moral theory. Definition (3) goes in the right direction, but does not take into account that the psychological plausibility of actions may need to be analysed from a social perspective.

If we define feasibility from a purely individual perspective, we blend out all actual social conditions which are given for socially embedded agents. But it cannot be right to assume that each individual can perform the same actions regardless of the social circumstances. What is feasible for an individual depends on what other people around this individual do. Definitions (3) and (4) can differ in two ways: On the one hand, acts which are unfeasible when looking only at individual psychological constraints may be feasible for a group, especially when the group can coordinate its actions. On the other hand, the individual perspective on psychological constraints fails to reveal that social action can be more difficult than individual action, in particular in the case of social dilemmas. Therefore, definition (4) can be both narrower and wider than definition (3), and there is no strict ordering regarding the feasibility of Φ_3 and Φ_4.[11]

There are at least four reasons why a dynamic, social perspective is preferable to an individualistic one. Firstly, humans often conditionalise their behaviour on other people's behaviour. For example, there is convincing evidence that people cooperate in PDs as long as they expect their opponents to cooperate (Gächter, 2006). Secondly, it seems that humans sometimes act as joint agents, rather than individual agents. What is implausible as an individual action might become plausible when agents apply "team thinking" and maximise the utility of the group (Bacharach, 2006). Thirdly, the interactions of moral agents are dynamic rather than static. What might seem unfeasible from a static perspective may be feasible from a dynamic one (and the other way round). Fourthly, agents act in pluralistic environments where different agents follow different norms. The effects of this pluralism can only be examined with a dynamic analysis. For these four reasons I regard definition (4) as the best definition of feasibility. Large parts of this thesis are concerned with the analysis of dynamic social interactions, and I claim that these explorations enhance our understanding of what is feasible moral behaviour for human agents.

[11] Also, definition (4) contains definition (3) as a special case, which makes it more general.

Both the descriptive and the normative perspective are better served with definition (4). An analysis that takes the social and dynamic character of constraints into account is descriptively superior because what is feasible for agents depends on how other agents behave. Definition (4) is also more convincing from a normative perspective because what one should morally demand from moral agents depends on the social circumstances they are in. Not all actions that look feasible from a physical or purely individual perspective are realistically possible for socially embedded agents. The upshot is that the set of feasible worlds Φ depends on the actual world α and the dynamic interactions of agents, where the agents are conceptualised with plausible psychological assumptions and where a plurality of psychological dispositions is taken into account.

What is feasible in realistic moral theory depends, among other factors, on the reinforcement mechanisms available. If compliance with a prescription can be enforced, the prescription is likely to be realistic. When compliance cannot be enforced, this is an indication that the norm might not be realistic. This can be demonstrated by looking at the social dilemma of norm compliance. If a norm that prescribes costly actions cannot be enforced, free-riding is the dominant strategy for payoff maximising agents. Complying agents are challenged by free riders, because free riders do better than cooperators. This is likely to lead to a downward spiral for cooperation. Obviously, many norms are obeyed, even though they prescribe costly actions, and formal policing and sanctioning is difficult. Therefore, we have to understand how the enforcement of these norms works. To do this, I analyse the social dynamics of norm compliance and norm enforcement. The focus is on "subtle" or informal mechanisms to enforce compliance. I argue that subtle sanctions are crucial for understanding the level of compliance with social norms observed in societies. Moral theory and social science have to work hand in hand at this point. In the following section I will look at one example of subtle mechanisms of norm enforcement.

3.4. Norm-Compliance and Subtle Mechanisms of Enforcement

In the introduction of this thesis I defined *social norms of cooperation* as norms that prescribe actions, which, when obeyed by all participating parties, produce greater average individual payoff, compared to settings where they are not in place and agents pursue individual utility maximisation. If all individuals pursue their own maximum payoff, everyone is worse off. The paradigmatic case is a public goods problem which

is equivalent to a n-person PD. In a public goods problem, the public good is funded by individual contributions. There are n individuals. The contribution of individual i is c_i. Each individual contributes either 0 or 1, which implies that, for each i, $c_i \in \{0, 1\}$.[12] The individual contributions are collected in a group fund, which is the sum of individual contributions $\sum_{i=1}^{n} c_i$. A collective investment is k times more profitable than an individual investment of the money. The problem is that the public good benefits everyone, whether they have contributed or not. Therefore, the payoff function P_i for individual i is

$$P_i = \frac{k \sum_{j=1}^{n} c_j}{n} - c_i.$$

With $1 < k < n$ it is more beneficial for an individual i to free-ride ($c_i = 0$) and benefit from whatever is contributed and shared by all other individuals. Since all rational, payoff-maximising individuals reason in the same way, no one contributes and the provision of the public good fails, although it would have been collectively beneficial.

Complying with norms of cooperation is a public goods problem if agents are pure payoff-maximisers. Still, all individuals should have an interest in maintaining and enforcing cooperation to reap the collective benefits from cooperation. In fact, a social norm of cooperation expresses the desire to overcome the collective action problem and therefore prescribes cooperation. The question is how a payoff maximiser can be convinced to cooperate. Clearly, there must be some form of side payment not included in the original statement of the public goods dilemma.

Let there be a side payment s_i that individual i has to pay if and only if i defects. If i contributes to the public good, the side payment is zero. In addition, there is a reward for any individual i who cooperates. Call the reward r_i. If the individual defects the reward is zero. When including side payments and rewards into the payoff function and distinguishing between cooperators and defectors, it changes to

[12] Using a discrete contribution function and a clear distinction between contribution and defection (also called free-riding) keeps the argument simple. With some additional formal effort one could replace the step function with continous contributions.

$$P_i = \begin{cases} \frac{k\sum_{j=1}^{n} c_j}{n} - c_i + r_i & \text{for cooperators } (c_i = 1) \\ \\ \frac{k\sum_{j=1}^{n} c_j}{n} - s_i & \text{for defectors } (c_i = 0). \end{cases}$$

A payoff-maximising individual needs to compare costs and benefits of defection. Call the difference of benefits and costs of defection, compared to cooperation, Δ. If Δ is positive, defection has a higher payoff than cooperation. When switching from cooperation to defection, an agent saves the contribution 1. At the same time, the agent loses r_i, has to pay s_i and also reduces the payoff from the public goods game by k/n. Therefore, we can compute

$$\Delta = 1 - r_i - s_i - \frac{k}{n}.$$

For large groups Δ approaches $1 - r_i - s_i$. The term $-r_i - s_i$ is the opportunity cost of defection created by the rewards for cooperation and the side payment for defection. Therefore, we can conclude that the opportunity costs of defection must be higher than the contribution of a cooperator to create an overall positive payoff for cooperators and to induce cooperation. Cooperation pays when the "punishment" of the side payments and the loss of rewards is higher than the gain from free-riding. When the opportunity costs of defection are high enough, the public goods dilemma is transformed into a cooperative game.[13] The question is now which forms the side payments and rewards might take. I argue that "subtle" forms of norm enforcement play an important role.

One possible way to implement side payments for norm transgression is a system of sanctions. Whenever a person violates the relevant norm, the person is punished. This punishment could be a fine, imprisonment, bodily harm, etc. It is well-known that punishment systems are difficult to implement because punishment and policing are costly. A working policing and sanctioning system is itself a public good, and producing it poses a second-order public goods dilemma. Sometimes the second-order problem can be solved. Often, however, sanctioning and policing are difficult to provide, in particular when policing is costly in comparison to the value of the public good. Public littering is a well-examined example (see Reno, Cialdini and Kallgren, 1993, for an empirical study) where the effort to police and punish people is much too high. Only

[13] A more realistic model would incorporate different utility functions for side payments. This would imply that some agents cooperate while others defect.

a dictatorial police state would be able to constantly monitor its citizens to detect and punish those who litter.

If people only complied with norms that are enforceable through formal punishment, social interaction would certainly look different. There must be other reasons why people obey social norms, in particular social norms of cooperation, even when they cannot be enforced or are not even codified as formal laws. Legal theorists have emphasised the importance of social norms, in contrast to legal norms (Ellickson, 1998; McAdams, 1997). A legal system may be built around enforced legal norms, but it would break down without many non-legal social norms that are obeyed by most people most of the time. Returning to the issue of side payments and rewards, I will now discuss several "subtle" sanctions and rewards to enforce norms without policing and sanctioning systems.

Reciprocity and Reputation. The mechanisms of direct and indirect reciprocity are promiment in the biological (see Nowak and Sigmund, 2005) and economic literature (see Sethi and Somanathan, 2003) on cooperation. eBay provides an example of how a system of "indirect reciprocity" (the technical term in the biological literature) works: For each transaction, buyer and seller rate the behaviour of their counterpart. These evaluations are stored and the track record of all eBay clients is publicly available. It is important to have a positive track record to find partners for future interactions. Having a positive reputation is an asset for a regular eBay trader. Therefore, it (usually) does not pay to cheat on the trading partner, even though it would be tempting to do so, given that the legal options are often limited, especially for cross-border trade with low-value products (who tries to proceed against someone in Taiwan over 60 Pounds?). The dynamics of indirect reciprocity can lead to intricate and sometimes surprising insights, as mathematical biologists have shown. In brief, systems of indirect reciprocity are prone to collapse under evolutionary dynamics, unless there are additional mechanisms to stabilise the norm. Nevertheless, examples of indirect reciprocity do exist and lead to high levels of cooperation, at least in the short and medium term.

Exclusion and Opting-Out. When cooperative individuals have a chance to exclude free-riders, or to opt out of the game when too many free-riders are around, then cooperation may be maintained. The effect is particularly strong when the game is played repeatedly and the track record of players is public knowledge. The more cooperative players will try to avoid free-riders. Those who consider free-riding will take this into account and may find that cooperation is overall more beneficial

when taking possible future exclusion into account. Exclusion and opting-out are related with the concept of reputation, as discussed above. A bad reputation can lead to exclusion. The reputation-tracking is the informational part of the sanctioning system, ostracism the actual sanction.

Desire for Esteem and Social Approval. These subtle sanctions differ from reciprocity and exclusion. The latter mechanisms are based on the "shadow of the future", which motivates agents to behave cooperatively in order to avoid future losses. The desire for esteem and social approval, by contrast, are psychological mechanisms with direct effects. Agents do not maximise their overall future payoff. Rather, they maximise their utility, *including their psychological utility.* These mechanisms work when people care about how other people think of them. If free-riding leads to a negative "image", and if agents care about their "image", cooperation may be the preferred choice because the psychological costs outweigh the material benefit of free-riding. Brennan and Pettit (2004) have formalised this mechanism and coined the term "Economy of Esteem" to analyse how the desire for esteem can lead to higher levels of compliance. Esteem can solve the second-order public goods dilemma of punishment because holding someone in esteem (or withdrawing the esteem) is costless.

Exclusion mechanisms play a crucial role in the translucency model of chapter 6. Reputation, reciprocity and ostracism also are important for the network models developed in chapter 7.

I have not mentioned another important "mechanism": The intrinsic motivation to do "the right thing". Undoubtedly, people often cooperate and follow norms because of their moral convictions. A Kantian might see cooperation as her duty, a utilitarian might think that he should comply with norms of cooperation because the collective utility increases. Denying these intrinsic or other-regarding motivations would be foolish. I do not rule out that they play a role, sometimes a crucial one. However, they are usually restrained by maximisation considerations when agents are threatened in their well-being.

Let me take stock of where we stand now. I want to develop ingredients to a realistic, positive theory of moral constraints. I have argued that a realistic theory must be based on a suitable notion of feasibility. This notion of feasibility must be a *social* notion. What is feasible for agents depends on the dynamic interaction between agents over time. I then moved on to the more specific issues of compliance with norms of cooperation, arguing that subtle enforcement mechanisms are crucial for understanding why

norms of cooperation are often obeyed, even though free-riding provides higher payoffs to each agent individually. The subtle mechanisms I have mentioned are all of a social nature, and a dynamic analysis of agent interactions is necessary for understanding these mechanisms.

However, the argument lacks one crucial step: I have silently assumed that free-riding will occur when an opportunity for free-riding arises. Moreover, I have claimed that the opportunity of free-riding kicks off a downward spiral for cooperation, with cooperation deterioriating as more and more people defect. Thus, I have smuggled in two anthropological assumptions: Firstly, I have assumed that agents desire material payoffs strongly enough to free ride when it pays; and secondly, I have assumed a dynamic by which "free-riding begets more free-riding over time", i. e. where agents tend to become more selfish when they experience selfish behaviour. So far I have not justified these two assumptions. One possible way to fill in this gap is to set up an evolutionary argument. To put this argument in a nutshell: Agents maximise payoffs because they face an evolutionary pressure to do so. The evolutionary argument also explains the dynamics: Over time, agents adapt the free-riding strategy because free-riding strategies promise a higher individual adaptive fitness. I will spell out the evolutionary argument in greater detail in chapter 5. In the remainder of this chapter, however, I show that the argument can often proceed without evolutionary ingredients.

3.5. Towards a Positive Theory of Moral Constraints

Moral theory is usually taken from the normative side, arguing what *should* be done. By contrast, this thesis focuses on what *can* be done by moral agents. I develop ingredients of a positive theory of moral constraints by using game theory and computer simulations. A positive theory of moral constraints is descriptive. Nevertheless, it informs prescriptive normative theory because moral theory for "men as they are" must take social reality into account.

Let me quickly rehearse the argument presented in this chapter. I began by defining realistic moral theory. To do this, it was necessary to find an adequate definition of feasibility. I argued that what is feasible for moral agents depends on their social interactions. Therefore, feasibility cannot be determined by merely looking at individual feasibility constraints. More specifically, I argued that compliance with norms of cooperation depends on subtle enforcement mechanisms driven by social dynamics. Many norms of cooperation would not be obeyed if there were no subtle social sanctions in place to make compliance the better choice at the bottom line.

Then I noted that my discussion regarding the dynamics of norm compliance depends on two anthropological assumptions in need of justification:

A1 Agents are at least partly driven by self-regarding utility maximisation.

A2 When agents see that other people behave selfishly, they tend to switch to a selfish strategy themselves.

The literature on the emergence of cooperation has often implicitly justified these two assumptions by looking at cooperation from an evolutionary perspective. If the selfish, free-riding strategy leads to higher adaptive fitness than the cooperative, other-regarding one, then the cooperative strategy is likely to be driven out in an evolutionary process.[14] The basic logic is that more selfish strategies lead to higher fitness, which leads in turn to higher replication. In terms of cultural evolution, individuals with selfish strategies are more successful, therefore more and more individuals imitate this successful strategy and play selfish, too.[15]

In my models of realistic moral constraints I will work mostly without evolutionary dynamics because less demanding, well supported assumptions about human psychology lead to models that are easier to justify. Therefore, I need a different justification for assumptions A1 and A2. Fortunately, there is plenty of empirical evidence to support these two assumptions. Lab experiments conducted by behavioural economists and social psychologists strongly support both A1 and A2. I do not deny that the ultimate causes for A1 and A2 may be identified by evolutionary theory, but for the purposes of my work I do not need to utilise contested evolutionary explanations because empirical evidence provides independent support for my assumptions.

Regarding A1, hundreds of experiments have shown that a majority of people tend to make payoff-maximising choices in dilemma situations (see Ledyard, 1994, for a review). The evidence clearly shows that people are at least partly acting as *homines economici.* Admittedly, individual payoff maximisation is by far not the only motive agents have, but this is not what A1 claims. The claim is only that economic considerations play an important role. One way to think about this is Philip Pettit's (2000) theory of

[14] I use guarded language here because the devil is in the details. A huge pile of literature regarding the evolutionary possibility of cooperation discusses settings where cooperation can be maintained even though cooperation is not directly fitness enhancing. But this should not conceal the fact that an evolutionary process usually leads to a breakdown of cooperation in a dilemma situation such as an n-person PD.

[15] Note that this argument does not work if group selection takes places. Cooperative groups do better *as groups*, compared to non-cooperative ones. This is one reason why some authors interested in showing the possibility of altruism try to revive the theory of group selection. See Sober and Wilson (1998).

"virtual self-regard" and his idea to analyse the "resilience" of behaviour. Put briefly, agents have many desires that are not payoff maximising. However, when essential material resources are under threat, payoff maximisation kicks in. This is why payoff maximisation is a resilient behaviour, even though it is usually only latently present. I discuss Pettit's view in greater detail in chapter 6.

Regarding A2, Fehr, Gächter and Fischbacher have found convincing evidence that many people are "conditional cooperators", that is, they cooperate as long as they expect others to cooperate too (Fehr and Fischbacher, 2004; Fischbacher and Gächter, 2006, with further references). It is indeed plausible that many people start with a willingness to cooperate, which supersedes payoff maximising considerations. However, when people feel that they are exploited because too many other people free-ride, they change to a less cooperative strategy. This is also indirectly supported by experiments showing that cooperation can be better maintained when agents have the option to exclude or avoid free riders (Hirshleifer and Rasmusen, 1989; de Vos, Smaniotta and Elsas, 2001; Cinyabuguma, Page and Putterman, 2005; Page, Putterman and Unel, 2005; Ones and Putterman, 2007).

Since A1 and A2 are well supported empirically and are often sufficient to set up models to explore moral constraints in realistic social settings, I do not see a need to use evolutionary models unless one has good reasons to assume that an evolutionary dynamic does indeed take place. Not all social dynamics are evolutionary dynamics. In chapters 6 and 7 will scrutinise social dynamics and I will make empirically supported assumptions about human behaviour along the lines of A1 and A2, but I will only make economical use of evolutionary arguments and models in this thesis. What matters to me are social dynamics of norm compliance and norm enforcement. These processes may have evolutionary aspects, but often they do not.

3.6. Conclusion

This thesis aims to provide elements of a realistic, descriptive theory of morality and social norms. In the first section of this chapter I provided essential conceptual ingredients to this project. To explore the realism of moral theories I have defined the concepts of feasibility, moral realism, and moral utopianism by using a simple possible worlds semantic. I have argued that a world is in the set of feasible worlds if it is proximate to the actual world. The possible transformations from the actual world to other feasible worlds are best analysed as a social, dynamic process. Therefore, dynamic agent-based models are an adequate tool to explore moral feasibility.

3. Subtle Ways of Norm Enforcement: Towards a Positive Theory of Moral Constraints

I have then turned to norm compliance as a special problem of feasibility. If we assume that agents are partly selfish and conditionalise their willingness to cooperate on the willingness of others to cooperate, then the compliance with norms of cooperation is threatened by defection, causing a downward spiral and a breakdown of cooperation. Thus, if compliance with norms of cooperation is supposed to be a realistic prescription, this downward spiral must be stopped by some form of social sanctioning. I will explore some sanctioning dynamics in the second part of this thesis.

4. Computational Models in the Social Sciences and Philosophy of Science

While the use of computational simulations is becoming more popular in the social sciences and in philosophy, there is little debate on the philosophical questions arising from the use of simulations to model social phenomena. It seems that, one the one hand, the enthusiastic embrace of these new methods has undercut the desire to understand the status of simulations in the explanation of social processes. On the other hand, skeptics have mostly ignored computational simulations as a new research methodology. This chapter aims to give a proper philosophy-of-science foundation for the use of social simulations, before practically applying simulations in chapters 6 and 7.

In this chapter, I argue for three positions:

1. Models in the social sciences should usually be highly idealised, given the epistemic and systematic restrictions faced when modelling complex social interactions and the resulting underdetermination of the model by the data.

2. The explanatory ambition of models in the social sciences is usually to find and analyse credible social mechanisms, rather than specific behaviours or histories in the actual world.

3. It is usually preferable to use small models with few variables and parameters rather than large models with many variables and parameters.

To argue for these three points, I begin, firstly, by describing the problem of representation, that is the problem of defining how model and target system are related. It turns out that many highly idealised models (HIM) make, strictly speaking, "false" assumptions regarding their target system. Nevertheless, HIM are useful because they have structural similarities with the target system. In the second section, I distinguish between two processes of idealisation: isolation and abstraction. I also propose an interpretation for the similarity relation between HIM and their target systems, in particular with regard to the analysis of social mechanisms. The third section introduces a

taxonomy for modelling approaches in the social sciences along the dimensions idealised vs. high fidelity and thought experiment vs. empirically grounded. I discuss several examples and point out that models in the social sciences are often underdetermined by the available data, in particular if researchers aim to model specific social processes. Finally, I argue that HIM in the social sciences should be small models because the plausibility of HIM hinges on their internal coherence and on the empirical support for their assumptions. If there are too many assumptions, such models becomes untestable and hard to analyse.

4.1. The Problem of Representation

Scientists are interested in models because models are supposed to represent something about the real world. Unfortunately, the representation relation is difficult to conceptualise, and has created headaches for philosophers of science. Frigg (2006) calls this problem the "enigma of representation" and asks:

> "Models are representations of a selected part or aspect of the world [...]. But in virtue of what is a model a representation of something else?" (Frigg, 2006, p. 50)

Before setting out my argument, I have to introduce some concepts. I use the terminology proposed by Weisberg (2003), which is largely similar to Giere's (1988). Figure 4.1 shows the relations between model description, model, and target system ("world" in Giere's terms).

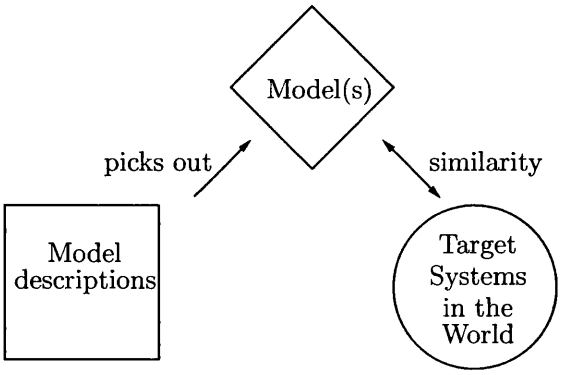

Figure 4.1.: Weisberg's concept of model individuation after Giere.

A *model description* consists of a system of equations or a program code. The model description has a number of *variables* and *parameters.* Weisberg calls a model description with fixed parameters an *instantiated model description.*[1] Usually we are interested in the change of variables over time, given certain parameter constellations. The change of the variables over time can be imagined as a trajectory through *state space.* The state space has as many dimensions as the model description has variables.[2] Each point in state space represents one unique state the model can be in. A *model* is one specific trajectory through state space as determined by the model description and fixed paramater values, according to Weisberg.[3] In contrast to Weisberg, I assume that a model can also be a family of trajectories through state space. Many models are stochastic and do not single out one specific trajectory. In these cases it would be counterintuitive to speak of distinct models.

Models relate to the world through a similarity relation. The models must bear some resemblance to the world to be of any relevance. In some cases this is easy to see. A toy train is a rather obvious model (a so-called scale model) of a real train. It has in many respects the same design as a real train, and to a certain extent it behaves like a real train. In many other respects it is not like a real train, of course, but it should be clear that toy train and real train are similar. Things are not that easy when model descriptions are equations or computer programs.

The similarity relation must be such that the model is relevant for the target system. Not any arbitrary similarity of instantiated properties will do. Unfortunately, the notion of relevance is difficult to pin down in a satisfying way. Teller (2001) may be right in claiming that "what is going to count as a relevant similarity depends on the details of the case at hand. No general account is needed precisely because it is the specifics of any case at hand which provides the basis for saying what counts as relevant similarity" (p. 401). I concur with Teller that a general account of relevance in model building may be difficult or perhaps even impossible. What demarcates relevant from irrelevant similarity is highly context sensitive.

[1] It is usually difficult to count the number of parameters in a model because many parameters are hidden as implicit modelling assumptions. It is good practice to make modelling assumptions as explicit as possible.

[2] This is true under the assumption that all variables are independent. Otherwise the state space can be smaller.

[3] This implies that two models can be identical even though they model different systems. For instance, the replicator dynamics can be a model for biological or cultural evolution. I still concur with Weisberg that the two models are identical. What differs is the similarity relation between target system and model. Thus, the same model can be used for different target systems.

4. Computational Models in the Social Sciences and Philosophy of Science

The representation question has been debated intensively for models in physics, and to a certain extent also for other natural sciences. However, virtually no work has been done to solve the problem of representation for models in the social sciences. This is regrettable because the problem of representation tends to be even more pressing, but also more difficult for the social sciences. In the social sciences, the gap between the target system and the model is usually wider than for the natural sciences. There are a number of reasons for that:

1. Social systems are often driven by unobservables (intentions, desires, beliefs, etc.).
2. Even if parameters and variables are in principle observable, it is often difficult to acquire valid and reliable data in sufficient quantities.
3. Experiments with real subjects are difficult or impossible.
4. Human decision making is not well understood and difficult to predict.
5. Social systems tend to be complex: They usually include distributed information processing, emergent properties, diversity of roles, multiple levels of interaction, and path dependencies.
6. It is often hard to identify one specific social process and treat it separately.

This means that models in the social sciences are difficult to build (because we often fail to simplify the complexity of reality into a sufficiently simple model) and difficult to test (because observations are difficult and the isolation of specific processes is often impossible).[4] Nevertheless, the use of models in the social sciences is possible and useful. Typically, models in the social sciences are highly idealised and work on a high level of generality. They are "thin" models (Kliemt, 1996).[5] In these cases, simulations are used primarily as exploratory devices. The researcher starts with some fairly simple assumptions for agent behaviour and uses the simulation for "if–then reasoning". This if–then reasoning can later be extended towards an analysis of social mechanisms, as explained below. What is necessary, however, is a better understanding of how these models relate to their target systems.

[4] However, this does not mean that model building in the natural sciences is easy. Unobservables, empirical challenges, complexity, and model identification are also difficult challenges when building, for instance, models in physics. See Cartwright (1999), in particular chapter 2.

[5] John Matthewson (personal communication) pointed out to me that there may also be other, non-pragmatic reasons for using simple models: Even if we understand a target system well, it may be preferable to simplify in order to focus on fundamental properties of the target system. However, in this chapter I mainly argue that simplicity is the right strategy for pragmatic reasons.

While looking for answers to these questions I found that economics tends to have better theoretical foundations and a livelier methodological debate than many other social science disciplines. This is probably because economists have been using abstract mathematical models to describe social behaviour for much longer than most other disciplines. The economic theorist Robert Sugden (2000) has tried to pin down the relation between abstract models and reality by discussing two examples: Akerlof's "Market for Lemons" (1970) and Schelling's "Segregation Model" (1978, p. 147–55) Here I will mainly focus on the latter. After giving a brief description of Schelling's segregation model, I discuss Sugden's view in detail and explain my own, differing perspective.

The beauty of Schelling's model is its simplicity. The model is set up by putting a number of pennies and dimes on a checkerboard. Some fields should be left free. Now you "run" the model by looking at each coin (the order does not really matter, as long as you make sure that every coin is looked at in the procedure) and determine how many coins of its own type and the other type are in its neighbourhood (take as neighbourhood all eight fields adjacent). A coin moves to a randomly determined empty spot when the coin has less than x percent of its own kind (penny or dime) in his neighbourhood. You could start with $x = 50\%$, but other values are possible. This moving rule is applied very often, until the checkerboard settles into a state where all agents are satisfied, i. e. no one wants to move. It turns out that the model almost always results in complete segregation, even for low values of x. A possible interpretation is that relatively weak individual preferences for "being with one's own kind" produce total segregation, even though agents individually do not require total segregation to be satisfied.

The purpose of Akerlof's model is to explain the trade-inhibiting effects of information asymmetries. To explain this effect, Akerlof discusses a highly stylised car market: "Suppose (for the sake of clarity rather than reality) that there are just four kinds of cars. There are new cars and used cars. There are good cars and bad cars (which in America are known as "lemons")." (Akerlof, 1970, p. 489). Now the market for used cars suffers from an information asymmetry: The seller knows whether his car is a good car or a "lemon". But the buyer cannot distinguish between good and bad cars. Therefore, good and bad used cars sell for the same price. This can lead to a situation where only "lemons" are offered because selling is attractive for the owners of "lemons", but not for the owners of good cars. The good cars are driven out of the market.

Akerlof's and Schelling's models are by now famous and almost universally recognised as outstanding theoretical works in economics. Regarding Akerlof, Sugden observes: "Akerlof has not proposed any hypothesis in a form that could be tested against ob-

servation. All he has presented is an empirically ill-defined "lemons-principle"" (p. 6). What we seem to get from Akerlof's paper is a highly unrealistic model of a fictional car market. Analogously, Schelling does not talk about real-world phenomena, but about cities on checkerboards, again without offering a testable hypothesis. What is the point of these papers?

Sugden discusses several perspectives on the two papers, in an attempt to figure out how Akerlof's and Schelling's abstract models relate to the "real world", or, more precisely, their respective target system. Firstly, Sugden asks whether the models are meant only as "conceptual explorations". If this was the case, Akerlof and Schelling would only be interested in discussing the internal consistency of their models, without reference to real phenomena. However, Sugden rejects this interpretation because "Akerlof and Schelling both devote such a lot of space to the discussion of real-world phenomena" (p. 10). Sugden then moves on to consider an instrumentalist interpretation of the models. He quickly rejects this thought on the grounds that Akerlof and Schelling do not offer any directly testable hypotheses.

Sugden begins to close in on his preferred answer when he discusses Gibbard's and Varian's interpretation of economic models as caricatures. The idea is that abstract models can be used to describe the world "in isolated or exaggerated form" (p. 14). Sugden elaborates on the idea of isolation and *ceteris paribus* claims by referring to the work of Uskali Mäki (1992) and Daniel Hausman (1992). Since the reality on the ground is messy, the economist (or the social scientist more generally) must isolate specific factors and analyse them separately. Models are "*thought experiments*" (Sugden, 2000, p. 15) to isolate and understand relevant factors. But this is still not a satisfying account of what Akerlof and Schelling do:

> "The difficulty for a Hausman-like or Mäki-like interpretation is that Akerlof's and Schelling's models both include many assumptions which neither are well-founded generalizations nor correspond with ceteris paribus or non-interference clauses in the empirical hypothesis that the modeller is advancing.
>
> [...]
>
> Take the case of Schelling's model. Suppose we read Schelling as claiming that *if* people lived in checkerboard cities, and *if* people came in just two colours, and *if* each person was content provided that at least a third of his neighbours were the same colour as him, and *if* ..., and *if* ... (going on to list all the properties of the model), *then* cities would be racially segregated.

> That is not an empirical claim at all: it is a theorem." (Sugden, 2000, p. 17)

The upshot is that Akerlof and Schelling do not make any direct claims about the actual world, but about counterfactual worlds. Still, these counterfactual worlds are not merely meant as thought experiments. According to Sugden, what Akerlof and Schelling give us are *credible worlds*. Following Sudgen, Schelling constructs a "set of *imaginary* cities" (p. 24). He shows that segregation appears in these imaginary cities as long as people dislike to be in a small minority. "We are invited to make the inductive reference that similar causal processes apply in real multi-ethnic cities" (p. 24). But why should we be willing to make this inductive leap of faith? Sugden argues that we are likely to accept Schelling's transfer from imaginary to real cities because "the model world *could be* real—that it describes a state of affairs that is *credible*" (p. 25). He links this concept of credibility to the credibility of characters and locations in novels. A model can describe imaginary, "false" events, but still be credible in the sense that these events could have happened. At this point, Sugden is close to Peter Godfrey-Smith (2006), who also argues that models often take the form of imagined realities, using the analogy with "literary fiction" (p. 735). But Godfrey-Smith is more careful: He only claims that scientists often treat their models like literary fiction, he does not claim that this is the right attitude towards abstract models.

I think that Sugden has this point wrong, and Godfrey-Smith is not sufficiently clear about it either. Literary fiction is credible when its authors get the details right. By contrast, models are good when their creators forget about the details and get the big picture right. When Henry James (1986), for example, describes "Daisy Miller", it is crucial for the credibility and beauty of his novella that he gets the details right about her American manners, the way she looks, and the way she talks ("She had ever so many friends who had been there ever so many times.", p. 56). When Schelling describes his checkerboard city, by contrast, it is crucial that he omits everything but the bare bones of the story to convince us that this model applies very generally. Schelling's model does not give us a credible account of real people in a real city. No city is like Schelling's checkerboard city. The criteria for credible fiction and credible models are very different from each other.

To get the relation between model and reality right, we have to return to Weisberg's and Giere's account, as depicted in figure 4.1. The aim is to describe the similarity relation between the model and the target system. I have argued that this similarity is not the same as that between literary fiction and reality. But what is it?

4.2. Identifying Social Mechanisms: Isolation and Generalisation

Models in the social sciences are usually highly idealised because they have to pick out very specific aspects of a complex and "messy" social reality. Given our inability to understand reality in all its complexity, we are forced to simplify and focus our research efforts on smaller entities where explanation is possible. These smaller entities are typically "social mechanisms". A social mechanism aims to describe a causal relation between some form of input and output in a social system:

> "Assume that we have observed a systematic relationship between two entities, say I and O. In order to explain the relationship between them we search for a mechanism, M, which is such that on occurence of the cause or input, I, it generates the effect or outcome, O." (Hedström and Swedberg, 1998, p. 7)

The point of the social mechanism is to open the "black box" and try to account for the reasons why a relationship between I and O exists. A social mechanism should be "robust" in the sense that it functions in many different environments. It should also be functionally independent in the sense that it can work without other social mechanisms and can be described independently from other social mechanisms.[6]

Social mechanisms are related to the discussion about mechanistic explanation in the sciences (see Machamer, Darden and Craver, 2000, for a review). As Glennan (2002) points out, mechanisms are particularly useful to analyse complex systems:

> "These mechanisms are systems consisting of stable arrangements of parts. In virtue of these arrangements, the systems as a whole have stable dispositions—the behaviors of these mechanisms. These dispositions can manifest themselves at more than one time and place. In this sense, the behavior of a complex-systems mechanism is general." (Glennan, 2002, p. S345)

[6] My "social mechanism" is in many ways similar to Woodward's (2002) definition of a mechanism. In particular, Woodward's definition emphasises the "invariance under interventions" (p. 370) a mechanism has to show. For example, if the mechanism alledges the generalisation G that a manipulation of X causes a change in Y, then the manipulation should work through X, and not through a different variable that is correlated with both X and Y. "Invariance" is given because G holds for at least some interventions on X (if it does not hold for any interventions, G cannot be a causal relationship). This implies that a mechanism must be "potentially usable for purposes of manipulation and control" (p. 370).

To analyse a mechanistic model, we need to meet two requirements according to Glennan: First, describe the mechanisms's behaviour. Second, describe the internal structure of the mechanism to account for this behaviour. This is similar to Hedström's and Svedberg's understanding of social mechanisms as attempts to open the "black box".

My use of the term "mechanism" should not be confused with Salmon's (1984) causal-mechanical model of explanation. Salmon's main worry is to develop a suitable concept of causality for scientific explanation. While this is important in its own right, it is not my ambition here. There are two orthogonal questions: First, what is the right notion of causality? Second, what is the right level of description for complex phenomena? I am concerned with the latter question. When describing a social mechanism, a researcher deliberately decides to undertake the explanatory enterprise on a high level. The mechanism itself consists of parts that stand in causal relations to each other, and these parts could be further reduced to more fundamental parts. Theoretically, if the problems of multiple realisation could be solved, this reduction could proceed until the mechanism is reduced to an explanation in terms of fundamental laws that cannot be reduced any further. At this level of fundamental laws, different notions of causality could be at work, but *for the analysis of the social mechanism it is not decisive which underlying concept of causality is used.* The point of the social mechanism approach is to explain by using high-level entities, namely the mechanism and its parts, because this level is most useful to explain the behaviour of the mechanism in its environment. I do not attempt to argue for a specific notion of causality. Here I differ from Glennan (2002), who relates his own approach to Salmon's causal-mechanical model. In my opinion, it is not helpful to conflate the question of the right descriptive level with the question of the right concept of causality.

Computational models, in my view, are a tool for identifying and analysing social mechanisms. To develop models of social mechanisms, modellers engage in two distinct processes of idealisation:

Isolation: The modeller picks out specific kinds of entities and their causal relations, and omits all other aspects of reality, to analyse the uninterfered and uninhibited effects of the hypothesised social mechanisms of interest.

Generalisation: The modeller assigns simplified quantitative and structural properties to the components of the model in order to represent a larger class of specific entities in the target system.[7]

[7]This list is not necessarily exhaustive. While I think that most idealisations are either isolations or generalisations, there might be other categories. For instance, sometimes researchers use deliberate

I explain both processes of idealisation by using Schelling's segregation model as an example. Schelling focuses on one specific social mechanism: segregation through repeated moves under the condition of weak preferences to be with one's own kind. He omits all other aspects that might be relevant for segregation: Politics, social norms, the housing market, schools, socio-economic processes, etc. By omitting all these potentially influential factors, Schelling considers only a very thin slice of reality to analyse the function of his hypothesised social mechanism, without interference from other factors (see Jones, 2005, p. 187).[8]

Schelling's social mechanism posits a causal relation between weak individual preferences for being with one's own kind and complete segregation, given an environment where agents move around when they are not satisfied with their neighbourhood. The model demonstrates the causal nature of the relationship by showing how these weak preferences and the rules for moving suffice to create segregation. It does so by *isolating* it from all other conceivable factors and mechanisms. One can demonstrate that the mechanism ceases to produce segregation if people have neutral preferences over whether their own kind should live in their neighborhood. Naturally, the model cannot prove that this is the actual mechanism taking place in reality. The model can only show that this mechanism can produce segregation. Whether this is the actual cause must ultimately be determined empirically.

Now I turn to the second aspect of idealisation: generalisation. I have argued that social science models only look at small slices of reality in order to isolate specific mechanisms. This is already a dramatic form of idealisation. However, usually this is not the only form of idealisation applied. After deciding to isolate a specific mechanism, the researcher typically *generalises* this mechanism to make her claim more universal. More precisely, after isolating the mechanism, the researcher looks at the remaining components in the model and tries to simplify quantitative and structural assumptions. The aim of this simplification is to capture a wide class of possible real world settings with a general model.

---

misrepresentation to improve the predictive success of their model. Misrepresentations may be a third form of idealisation. However, they are difficult to justify, especially when the target system is not well understood, unless they are either needed to isolate or generalise the model.

[8] Sugden is probably right that Schelling did not develop his model by starting with reality and gradually simplyfying it. Rather, Schelling probably started with a (rough) hypothesis, which must have looked like this: "Segregation occurs whenever people have weak preferences to be close to some of their own kind and people can move from less preferred to more preferred neighborhoods". He then probably began experimenting and came up with the checkerboard model. Therefore, Schelling did not start from the real world, reducing it to the model. Rather, he came up with a constructed HIM to look at the mechanism of interest under conditions of perfect isolation.

In Schelling's case, real agents have different numbers of neighbours, and they have different perceptions of whom they count as a neighbour. On the checkerboard, each agent has at most eight neighbours (with the exception of agents at the border of the checkerboard, but this could easily be avoided by "wrapping" the field and playing it on a torus). Assuming that each agent has eight or fewer neighbours does not seem like a generalisation. However, the important point is that it does not matter for the model how many neighbours an agent has. The mechanism described works for many different neighborhood topologies (see Fagiolo, Valente and Vriend, 2005). Schelling does not make this generalisation explicit, but it is clearly implied that his mechanism is meant as a general mechanism which is invariant under many changes to the actual topology. Further generalisations are made with reference to agents' preferences, the process of moving, etc. The point is that the mechanism works with many different parameter values, start values, and structural assumptions.

The advantage of general models is twofold. Firstly, we often do not know the target system very well. As explained above, the social sciences analyse social phenomena that are difficult to capture in all details. Therefore, it is useful to have a general model because we can argue for a similarity relation between model and target system even if our knowledge about the target system is limited. General models lend themselves more easily to the analysis of target systems that are not fully understood. Secondly, even if the target system is well understood, it is useful to have models that apply to different target systems. In the case of Schelling, we can think of many target systems in which his mechanism applies. Some of them are real systems, others could be real. Schelling himself suggests that his segregation mechanism might influence racial segregation in the housing market, or seating arrangements in sports arenas. It might apply to the clustering of men and women at parties, or to the segregation between rich and poor, to name just a few examples. The power of the mechanism lies in its generality.

I return to the orginal question: In virtue of what is a model like Schelling's a representation of a real target system? Asked differently: What is the similarity relation between model and target system? The analysis so far does not provide a full answer, but at least it points us in the right direction. The model is in many ways *not* similar to the target system because the modeller deliberately simplifies, omits, and generalises. Therefore, the relation between model and target system is not isomorphic.[9] Rather, it must be based on a *similarity of instantiated properties* (Teller, 2001). The similarity can be quantitative, qualitative, or structural. For example, in Schelling's model there

[9]Semantic theories of models assert isomorphism, but this view has come under severe attack. For a review see, e. g., Godfrey-Smith (2006) or Frigg (2006).

is a structural similarity between real neighbourhoods and the model neighbourhoods on the checkerboard. The model neighbourhood has the property that agents have neighbours, and that they care about the type of their neighbours. Also, the model has the property that agents can move. Potential target systems (say, the housing market in London) also instantiate these properties: People have neighbours, and people care about who they have as neighbours, and people can move if they do not like their neighbours. The model and the target system instantiate the same structural properties and are in this regard similar. This similarity with regard to some structural properties suffices to convince us that Schelling's model represents some important aspects of the target system.

Note that Schelling does not claim that his model gives a full account as to why we find segregation in many target systems. He only claims that his model describes a social mechanism that offers one possible explanation of segregation. The model isolates this mechanism and only this mechanism. Insofar as it omits and idealises it is not a "true" model of the target system. But it is a plausible model of reality with regard to the social mechanism because of its structural similarities with the target system: The social mechanism works whenever a social system has certain structural properties (neighbours, preferences over neighbours, the option of moving, etc.). Since many social systems do indeed instantiate these properties, it is plausible that the social mechanism is at work in these social systems.

To summarise: Models in the social sciences are typically highly idealised because it is difficult to capture all aspects of social systems precisely. Under these conditions, modellers focus on the discovery of social mechanisms. To identify and analyse social mechanisms, modellers apply two distinct idealisation techniques: isolation and generalisation. The resulting highly idealised models stand in a similarity relation with their target system insofar as they instantiate similar quantitative, qualitative, or structural properties as their target systems.

I have described the scientific practice of model builders in the social sciences and clarified concepts. I have also argued for isolation and generalisation as model building techniques to identify and analyse social mechanisms. In the next sections I focus on the question of how model building in the social sciences should be pursued and why I think that highly idealised, small models are usually the better choice.

4.3. Two Dimensions for Models

One can classify simulation models in the social sciences along the dimension of idealisation. On the one hand, there are *highly idealised models* such as Axelrod's (1984) computer tournaments or Skyrms's (1996; 2004) games with replicator dynamics. On the other, we find *high fidelity models* with many variables and parameters, aiming to model the development of concrete, real social processes. An example is Dean's et al. (2000) simulation of the Anasazi Cultural Change (see also Axtell et al., 2006). A second dimension to distinguish models is whether they are *thought experiments* or *empirically grounded experiments.* Figure 4.2 displays both dimensions. The two dimensions are correlated: Thought experiments tend to be idealised, empirically grounded models tend to aim at high fidelity. For instance, most of Axelrod's simulations fall into quadrant III, while Dean et al. are in quadrant I.

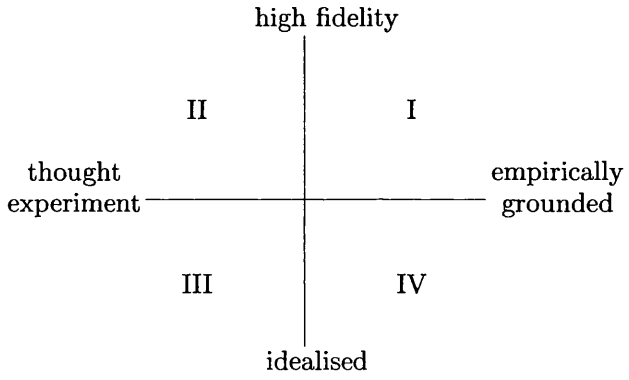

Figure 4.2.: Two dimensions of simulations in the social sciences.

I will argue that it is usually preferable for social science simulations to remain on a high level of idealisation. If possible, they should move from a mere thought experiment to empirically grounded simulations, i. e. from quadrant III to quadrant IV. High fidelity models, by contrast, should only be used in exceptional circumstances, and only if they rest on firm empirical grounds. Thus, I will argue that models in quadrant I only make sense for specific, well-defined problems, while models in quadrant II never make sense.

HIM usually have few variables and therefore a small state space. For example, take Skyrms's model of "cake sharing" (1996, ch. 1): For each interaction, two players meet at random and play the following game: The two players have to share a cake (assume the cake is "manna from heaven"). Both players specify how much of the cake they demand. Player A demands a percent, player B b percent. If the sum of the demands

is smaller than or equal to what is available ($a + b \leq 100$), the two players get what they demanded (and all leftovers are disposed). If the sum of the demands is greater than what is available, neither player gets any cake. In the model, players have fixed strategies, that is, they make the same demand all the time. Skyrms's model also contains a replicator dynamic: The relative frequency of a strategy in the population depends on its relative success in getting cake. Roughly speaking, players with more cake replicate fast, players with less cake replicate more slowly. Players without any cake die out. Skyrms then looks at the relative frequency of strategies in the population over time and determines the basins of attraction for stable equilibria (stable strategy proportions in the population).[10]

Skyrms's model description takes as parameters the initial state of the population, i.e. which strategies are present in which proportion. Further implicit parameters are given by the way the interaction is spelt out and how the replicator dynamics is implemented. The variables for this model are the proportions of these strategies in the population. Thus, running the instantiated model description produces a trajectory of these strategy proportions in the state space.[11]

Clearly, Skyrms is not really interested in the population dynamics of cake-eaters. What he is interested in is the "evolution of justice" (p. 20) and sharing norms. Note that his model does not rest on any empirical evidence, nor does Skyrms compare the results of his model with sharing behaviour in the real world (except in a very anecdotal manner, appealing to the reader's experience and intuition). Therefore, the model is positioned in quadrant III of figure 4.2.

I now contrast Skyrms's HIM with two different high fidelity models. The first is designed by the Redfish Group in Santa Fe. They scrutinise the dynamics of crowd movement in a sports stadium. More specifically, they simulate crowd movement after a bomb explodes inside the stadium. The aim of this simulation is to improve evacuation procedures and minimise casualties in case of an emergency.[12] This is a high fidelity model because it uses detailed geographic and environmental information and makes relatively complicated assumptions regarding agent behaviour. It is also, at least to a certain extent, an empirically grounded model. The geographic and environmental factors were recorded, and the agent behaviour can be based on observations of real agents

[10] A precise specification of the model should spell out whether the population is finite or infinite, and if the former, whether the number of agents is normalised after each application of the replicator dynamics.

[11] In an infinite population an instantiated model is deterministic and has exactly one trajectory. If the population is finite the random matching of strategies is a stochastic process and there are many (converging) trajectories.

[12] see `www.redfish.com/stadium`

in a sports stadium and can even be tested in a large scale experiment (a live evacuation simulation was undertaken at the Pittsburgh PNC Baseball Stadium). Thus, the modelling assumptions are based on observations and some reasonable assumptions. The model can also be tested against existing knowledge of crowd behaviour in emergency situations. Overall, there are good reasons to be confident that the model captures the dynamics of the crowd quite well and is able to produce realistic scenarios of what might happen. Naturally, the model cannot predict the precise course of events, but it is useful to get an idea of likely crowd behaviour.

The second model is the "Anasazi Culture Change Model" by Dean et al. (2000), with further additions in Axtell et al. (2006).[13] Dean and his collaborators use an agent-based model to reconstruct the "spatiotemporal history" (Dean et al., 2000, p. 180) of the Anasazi, who lived in the Long House Valley in northeastern Arizona between 1800 B. C. and A. D. 1300. The population dynamics and the agricultultural conditions in the valley are quite well known due to archeological and geological research. The idea is that the model can be tested against the available data:

> "The degree of fit between the model and real-world situations allows the model's validity to be assessed. A close fit between all or a part of a model and the test data indicates that the model, albeit highly simplified, has explanatory power." (Dean et al., 2000, p. 180)

Again, this model aims to be in quadrant I of figure 4.2.[14] However, in contrast to the Redfish stadium simulation, the empirical foundations are much more problematic in this case. The problems are partly empirical, partly theoretical.

Dean et al. argue that the population development predicted by their model is quite similar to the reconstructed real population development. However, this claim is questionable: Firstly, the "real" population development is unknown and the model can only be compared with estimates based on archeological data. Secondly, model and (estimated) reality deviate qualitatively and quantitatively, especially in Dean et al. (2000). While both curves show roughly the same local maxima and minima (p. 191), the model fails to predict the collapse of the Anasazi culture in the valley around A. D. 1300. Also, the model predicts the existence of around five times more households than the archeological data suggests. In addition, Dean et al. only show the average value

[13] This model is also discussed by Grüne-Yanoff (2006). He is mainly concerned with the question of explanation. I say more about his view below.

[14] Even though Dean et al. state that the model is "highly simplified" it is a high fidelity model in my sense because it aims to model an actual historical development, rather than just exploring an isolated social mechanism.

of 35 simulation runs (p. 190). It is not clear whether different runs would produce entirely different results. If yes, this would put the alleged fit between model and reality into question. Later, Axtell et al. (2006) extended and changed the model, producing a much better fit between estimated history and model prediction.[15] But even if the model fitted the data perfectly, this does not mean that the model is similar to the actual data generating process. In fact, the idea to add additional parameters just to increase the model fit is in itself a problem: One can always "torture the data until it confesses", and there will always be fitting models if one changes the model long enough. Alexander (2003*a*, section 4.2) notes that "nearly any result can be produced by a model by suitable adjusting of the dynamics and initial conditions". The decisive question is: Why is one specific model the "right" one, and not any other model that might fit the data equally well?

I do not criticise Dean et al. because I think this is a bad application of computational simulations in the social sciences. To the contrary: Dean and his collaborators are aware of the difficulties, and it is only natural that such an ambitious project faces problems. What I want to emphasise are the theoretical challenges when building empirically grounded high fidelity models in the social sciences. Put briefly, high fidelity models face the problem of overfitting. The Anasazi culture model has dozens of parameters, at least when counting all the implicit assumptions contained in the simulations. This creates a huge parameter space. Each point in this parameter space represents a model instantiation. At the same time, there is a fairly limited set of estimated data points describing the real development in the Long House Valley. Many, potentially infinitely many model instantiations in the parameter space might fit this limited set of data points quite well. Unfortunately many of these model instantiations do not have much similarity with the actual data generating process, that is the actual history of the Anasazi Culture. The bottom line is that the model is dramatically underdetermined by the data available.

In principle, almost every model in the social sciences faces the problem of underdetermination. However, the problem is mitigated in cases where the parameter assumptions can be tested by empirical work. Redfish's crowd dynamics rest on empirical knowledge on the micro-level (how do agents behave?) and on the macro-level (how do crowds behave?). Also, it is reasonable to assume that agents tend to apply simple heuristics

[15] Still, methodological questions remain. Axtell and his collaborators give us their "best single run" (p. 125), and they also claim that the average run fits well, without giving us data for the average run. This seems to suggest that there are many other runs that do not fit so well. Since the model is stochastic, it is not surprising that some runs fit well. It would be more interesting to see how many runs fit well and whether they all show structural similarities.

in stress situations and that heuristics like "look for the exit you used to enter the stadium" are likely to be good approximations of real agent behaviour. The Anasazi Culture Change Model, by contrast, deals with rich cultural phenomena. Social norms regulating tribe and family relations, sharing behaviour, traditions, etc. can fundamentally change the model. However, they cannot be accounted for in the model due to lack of information. Therefore, it is not clear whether the Anasazi Culture Change Model has much explanatory power when it comes to reconstructing the spatiotemporal history in the Long House Valley. Even if it fitted the data perfectly, it might do so for entirely wrong reasons.

Most problems in the social sciences are unlike the Redfish crowd model, but more like the Anasazi culture model. Often we simply do not know enough to justify the assumptions of high fidelity models empirically. Some researchers have reacted with a defiant "pioneer" spirit, i. e. they still build high fidelity models in the hope that our methods of verification will improve. Others try to develop auxiliary arguments as to why we have reasons to be optimistic about the future of high fidelity models. I discuss one of these arguments in chapter 5. A third group claims that small models or "thin simulations" (Kliemt, 1996) are the right methodological choice when our knowledge of the target system is limited. I argue for this last approach.

4.4. The Virtue of Small Models

The argument presented so far consists of two parts. In the first part, I have discussed the problem of representation. Models in the social sciences usually focus on very specific "slices" of reality. The reason for this is that social systems tend to be complex, while our knowledge of them is still very limited. Given these limitations, modellers should focus on specific social mechanisms, rather than on the system as a whole. I have shown that we need two distinct processes of idealisation to build models of social mechanisms: On the one hand, the researcher must isolate the social mechanism. On the other, the researcher must generalise to show that the hypothesised social mechanism applies in many credible or actual social systems. This leads to the claim that the model and the target system relate through a limited form of similarity, namely by instantiating similar properties in some specific respects.

The second part of the argument has explained why highly idealised models are preferable to high fidelity models in the social sciences. Modelling social science processes with "thick" simulations is questionable because there is usually not enough data available to validate the model's specific assumptions and settings. Therefore, modellers

in the social sciences are better off in restricting themselves to small, highly idealised models.

An additional reason for using small models is that—paradoxically—it is often easier to argue for the similarity between model and target system if the model is small. Adding more details to a model seems to increase the similarity between model and target system. However, with more details and a finer structure it becomes more difficult to show the alledged similarity. To see this imagine a more refined version of Schelling's segregation model: One could add details by using estimates of people's real preferences. Also, one could try to replace the unrealistic checkerboard with a more realistic topology. In this way one could work towards a more realistic model of a specific target system. The model is more similar to its specific target system now. But the causal story proposed by the model is now much less general.

It is easy to argue that Schelling's simple model is similar to its target systems because it only presupposes similarity regarding some very general structural properties. It is much harder to argue that the more detailed model is similar to its target system because it presupposes many more properties of the system. The problem with the more detailed model is that it can get the causal story wrong in many more ways than the simple model can. Also, the more detailed model does not add anything to the explanation provided by the simpler model. Therefore, when the knowledge of the target system is limited, restriction to simpler models is often preferable.

A similar point is made by Boyd and Richerson (2005, ch. 19). Arguing on the basis of their modelling practice in the field of cultural evolution, they dismiss high fidelity models "because (1) they are hard to understand, (2) they are difficult to analyze, and (3) they are often no more useful for prediction than simple models." (p. 402). They also call for a "modularization of analysis" (p. 406) similar to my strategy of analysing social mechanisms by isolation and generalisation. Boyd and Richerson emphasise that HIM are superior to high fidelity models not only because they are easier to validate, but also because they are easier to understand and are therefore better suited to explain complex phenomena. A more complex model might be more similar to the target system, but if we do not properly understand the complex model, we cannot even be sure whether it really is a better approximation of the target system.

The two parts of the argument presented in this chapter support each other. The second part argues for HIM from a practical perspective, the first part lays the theoretical foundations for HIM from a philosophy of science perspective. When put together, the two parts form an argument for a specific research paradigm regarding the use of computational models in the social sciences. This paradigm is based on social mechanisms

as the core theoretical element. It uses the methods of isolation and generalisation to define model descriptions and analyses the resulting model(s) by means of computational simulations. Therefore, it aims to explain causal processes in the target system by analysing causal mechanism in the controlled environment of the simulation.

Joshua Epstein and Robert Axtell argue that the purpose of computational models in the social sciences is "Growing Artifical Societies"—or at least this is the title of their 1996 book. I think this picture is misleading: It suggests that the modeller produces a society that is (though "artificial") similar to a real society. I have argued that this is hardly the case. In his "Generative Social Science" Epstein (2006) is more careful. He claims that "[i]f you didn't grow it, you didn't explain its emergence" (p. 8). However, he also notes that "if a microspecification, m, generates a macrostructure of interest, then m is a *candidate* explanation. But it may be a relatively weak candidate; merely generating a macrostructure does not necessarily explain its formation particularly well." (p. 9). The question is how to distinguish between different candidates. I concur with Epstein that this can only be done empirically.

There are two ways to add empirical credibility to a model: First, the model should produce plausible outcomes on the macro-level. If the aim is to model a concrete social system, then the model should match the behaviour of the target system closely. If the model models a social mechanism, structural similarities are more important. The second way to achieve empirical credibility consists in empirical support on the micro-level. This means that the assumptions and parameters of the model should be empirically supported.[16] For example, Schelling's model gains empirical credibility if it is supported by empirical data regarding the segregation preferences people have. The idea is to test whether the underlying assumptions of the model are empirically justified. If this is the case, the model moves from a mere thought experiment (quadrant III in figure 4.2) to an empirically grounded HIM.

Grüne-Yanoff (2006) points to a second problem with Epstein's view: Epstein claims that the model as such is the *explanans* of some emerging macrostructure (the *explanandum*). However, the similarity between model and target system is limited and often only of a structural nature. The model may offer an explanation of one possible social mechanism leading from some input to some output. But the target system consists of much more than this one social mechanism described by the model. Therefore, it is mistaken to assume that the model is the *explanans* of the target system. Rather,

[16]The current advances in behavioural economics may be particularly useful in this regard: Testing assumptions about agent behaviour is—at least to a certain extent—possible by using controlled experiments.

the model is the *explanans* for a different *explanandum*—the disposition of the target system to behave in a certain way (Grüne-Yanoff, 2006, p. 13).[17] This is consistent with my argument for small models: If the model aims to explain the dispositions of the target system, then it is better to find simple explanations for these dispositions, particularly if the model is difficult to validate.

4.5. Conclusion

In this chapter I have argued for the use of highly idealised models in the social sciences. The argument consists of two parts. I started with the problem of representation to show that the analysis of complex social systems is best conducted by looking for social mechanisms. These social mechanisms have a similarity relation with their target system insofar as they instantiate the same (often structural) properties as the target system with regard to one specific aspect of reality. Analysing social mechanisms involves strong idealisations. Isolation is used to exclude irrelevant factors and to focus on one specific mechanism; generalisation is used to make claims over a wide range of different but relevantly similar target systems. HIM do not aim to model the target system as a whole, they rather explain its dispositions.

In the second part of the argument I have developed a taxonomy of model types in the social sciences along the dimensions high fidelity vs. idealised and thought experiment vs. empirically grounded. The critique of high fidelity models has led to an argument for highly idealised models. A core problem for models in the social sciences is their validation: There can be many, potentially infinitely many models competing to explain the same outcomes. This problem is easier to address if the model is small and if the modeller focuses on specific social mechanisms.

In this thesis I develop two models. Both will be highly idealised accounts of social mechanisms that might induce cooperation in dilemma situations. Undoubtedly, the reasons why people cooperate, despite being in a social dilemma situation, are complex. In line with the argument put forward in this chapter, I do not attempt to model this complexity, and instead argue that my models capture important structural properties of real social dilemmas.

[17] In addition, Grüne-Yanoff also stresses the validation problem for models. If there is a potentially infinite number of models producing the same results, one does not explain very much by looking at merely one of these models. Epstein himself acknowledges that the unlimited number of options for the specification of the model is a great problem for the generative approach (Epstein, 2006, p. 30).

5. Some Remarks on Evolutionary Explanations of Norms

So much has been written on evolutionary theories of human behaviour that reviewing this literature alone could fill a book. In this chapter, my target is more limited: I concentrate on the question as to how and to what extent evolutionary explanations help to explain and possibly predict social norms. The attitudes towards the explanatory usefulness of evolutionary considerations range from wild optimism to deep scepticism. These differences are often rooted in scientific disagreements regarding the analysis of evolutionary processes or in philosophical disagreements about how evolutionary theories can be used as explanations. The debate is complicated because positive and normative arguments are often intertwined.

The restriction to "some remarks" is in order because the subject at hand is complex and the views are still changing rapidly as our knowledge of evolutionary theory expands. I argue for a skeptical position here, but empirical results may prove me wrong in the future. In this chapter I begin by giving a short and simplified account of biological and cultural evolution. Sections 2 and 3 discuss the pros and cons of biological and cultural evolution as explanatory tool for social behaviour, with special attention to social norms. Section 4 is an interlude, connecting evolutionary theories with issues of model building as discussed in chapter 4. Section 5 criticises "grand sociobiological theories" and argues for the application of small models of evolutionary game theory. Section 6 wraps up the argument, claiming that the most important function for evolutionary analyses of human social behaviour is the discovery of feasibility constraints, creating a link to chapter 3.

5.1. Biological and Cultural Evolution

With "biological evolution" I refer to the mainstream Darwinian theory of evolution. While Darwin was unaware of genes, nowadays genes are assumed to be the basic

entities of selection.[1] Dawkins's "selfish gene" and "vehicle" metaphor (1989), despite all the controversies it has caused, may be helpful for expressing the fundamental idea. Metaphorically speaking, a gene is interested in replicating itself. The success of the replication depends on the replicator's phenotypic effects, or its "vehicle". Human genes, for example, are in this metaphorical sense interested in replicating. How successfully one gene replicates depends on how well its "vehicle", one instantiation of *homo sapiens*, does in a given environment. The unit of selection is the gene, but the replicative success depends on the phenotype and the environment it operates in.

Evolution occurs because there are fitness differences between competing genes, and because variations of genes occur due to recombination and mutation. Natural selection tends to favour those genes that have phenotypic effects leading to higher adaptive fitness. Since recombination and mutation provide a constant source of "innovation", the existing genes compete with new candidate genes all the time. Sometimes a new gene is "better" in the sense that it makes its "vehicle" more successful in replicating the gene, compared to other genes. Then this new gene will replicate faster and will eventually crowd out its competitors.

This coarse-grained outline of biological evolution would need several qualifications to approach the sophistication of the current debate in evolutionary theory. This thesis is not the place for these qualifications. Since my aim in this chapter is to debate the explanatory power of evolutionary theory for social norms, it is more important to think about how biological evolution has shaped humans to become norm-guided beings. To promote this aim, one must note that *homo sapiens* is more or less biologically unchanged since its hunter-and-gatherer times in the *pleistocene*. Thus, the environment of evolutionary adaptedness (EEA) looked radically different from modern 21st century societies. As far as biological evolution is concerned, we should expect humans to be optimised for life in a troop of a few dozens of hunters and gatherers. I will discuss below what to make of this consideration regarding social norms.

A different, and more contested, form of evolution is of greater relevance for the explanation of social norms: cultural evolution. There is still no unified paradigm of how to think about cultural evolution. Speaking in very general terms, cultural evolution posits that pieces of cultural information (for example skills, artefacts, ideas, and perhaps norms), stored in human minds, are replicated. This replication is supposed to be similar to the replication of genes in genetical evolution. Replication[2] can hap-

[1] However, genes are most likely not the only level of selection, and some theorists doubt that the gene-centred view is helpful. See Lloyd (2005) for a review and further references.

[2] The term "replication" may be misleading because cultural pieces of information are usually not

pen through different mechanisms. Agents may imitate others, or they may actively learn from others. Imitation and learning differ: Imitation results in copying, while learning involves a higher level of cognitive processing, which often entails changes to the information input. It is also possible that information is stored in artefacts that can be reverse-engineered (Sterelny, 2006, p. 12–13). In addition, differences arise because information can either flow horizontally between agents of the same generation, or vertically when information is passed on from parents to their children. Depending on the flow of information, the level of selection can differ: If information is distributed horizontally, then all agents in the same social group engage in the same or very similar social practices, while other groups have different practices. Thus, for horizontal information transmission the evolutionary process must occur at the levels of groups. If, by contrast, the information flows vertically, then fitness differences occur at the level of the individual agents.[3]

Dawkins (1989) argues that "memes" play the role of genes for cultural evolution. According to this view, memes carry cultural information and replicate because memes leading to higher adaptive fitness are imitated more often. The question of how well the analogy between genes and memes works is subject to debate, and Dawkins himself treats it with caution. Other proponents of cultural evolution reject the idea of "memetics", arguing that the analogy is flawed (see contributions to Aunger, 2000). However, even if the analogy between memes and genes may be a red herring, cultural evolution can still work through different mechanisms. (Boyd and Richerson, 2005, ch. 20). For example, the models developed by Boyd and Richerson (1985, 2005) apply Darwinian thinking to cultural practices, without presupposing the existence of memes.

Dennett (2006, p. 341) remarks that evolution is "substrate-neutral" and "will occur whenever and wherever three conditions are met:

1. replication

2. variation (mutation)

copied with high fidelity. Still, a Darwinian process may occur, as Gil-White (2005) argues. For simplicity, I stick with the term replicator, but I do not deny that many processes of cultural evolution work with low fidelity inheritance. Exploring the precisce mechanisms of information transition is important to develop a full-fledged account of cultural evolution.

[3]The vertical flow of information is crucial when considering the evolution of norms because it is in the nature of norms to be publicly known. Sterelny (2006) sees the horizontal flow of information and group selection as core elements of "niche construction". The term "niche construction" alludes to the fact that human agents (or organisms more generally) change the environment they operate in for themselves and for future generations. Thereby they can create "niches", which allow them to increase their adaptive fitness. I will argue below that niche construction is the most important transmission mechanism for the inter-group competition of norms.

3. differential fitness (competition)".[4]

The substrate-neutral character of evolution suggests that it can be applied to culture. Dennett calls the units of selection "universal replicators". In his view, replication of cultural information is given by transmission mechanisms (learning, imitation, ...); variation is given by errors in transmission or attempts to try something new. But what about differential fitness? Dawkins and Dennett emphasise that the fitness of a "meme" or "universal replicator" is measured by the expected replicative success of the meme. It is *not* measured by the fitness of the agent applying or believing the content of the "meme". For instance, the idea to use hammers to hit nails is in competition with other ideas about how to drive a nail into a wall. The fitness of the "hammer idea" is measured in terms of how well we expect it to replicate, not in the fitness of people using a hammer.[5]

Dennett's substrate-neutral account of evolution works on the conceptual level. But to apply it to cultural phenomena, we need a much better understanding of how replication and variation work in reality. As mentioned, there are many different mechanisms of cultural information transmission, and not all of them presuppose the existence of replicators. The severe critique mounted against the "memetics" research programme suggests that the idea of universal replicators is contested.

5.2. Biological Evolution and Norms

What can moral philosophy learn from the theory of evolution and how does evolution have an impact on morality and social norms? To begin with, human agents are the product of biological evolution. At some point, the biological setup and the hardwired psychological mechanisms of humans must have enabled them to develop moral systems. I believe that this is largely uncontroversial. More controversial are claims that social norms are themselves the products of evolutionary processes.

Humans come with some remarkable abilities not found in other species. One of them is the ability to communicate in a highly complex language. Another ability is to empathise with other people. A third ability is a highly developed self-consciousness. It is uncontroversial to assume that genetic dispositions are foundational for these abilities. For instance, the fact that we have a system to produce complex sounds with our voice is biological, not cultural. Why did evolution produce animals with these abilities? The

[4]The observation that evolution applies to many "Darwinian populations" was first made by Lewontin (1970).

[5]This last point is contested, and I explain my position in greater detail below.

likely answer is that such animals had a higher adaptive fitness than animals without these abilities. It seems that the ability to develop a language, to empathise, and to have self-conciousness guaranteed our evolutionary success. Sociobiologists argue that this should already teach us something about our morality: Our morality is constrained by our biological setup, and the way we are biologically hard-wired was determined by an evolutionary process.[6]

Evolutionary psychologists try to answer the question of how we are biologically hardwired (Tooby and Cosmides, 1992). Put in a nutshell, they try to explain human behaviour by identifying psychological mechanisms (often called "modules" or "circuits in the brain") that evolved during the EEA and influence our behaviour today. Evolutionary psychology is not an attempt to reduce human behaviour to adapted mechanisms, but it does make the claim that aspects of behaviour are at least partly due to evolutionary causes, and therefore not completely "learned" or "cultural". Dennett, sympathetic to Tooby's and Cosmides's approach, nails the controversy neatly:

> "Even if they [Tooby and Cosmides] are right—and I am confident that they are—that such rationality as we human beings have is the product of the activities of a host of special-purpose gadgets designed by natural selection, it does not follow that this "Swiss-army knife" of ours cannot have been used, time and time again, to reinvent the wheel. It still has to be shown, in other words, that any particular adaptation is *not* a cultural product responding quite directly (and rationally) to quite recent conditions." (Dennett, 1996, p. 490)

Dennett points out that it is often difficult to distinguish between biological adaptations and cultural constructions. Humans are biologically constrained, but they are also amazingly good at overcoming these constraints and finding solutions to new problems, even though they are—as a species—constructed in the *pleistocene*. The hardwired "gadgets" in our brain that were adaptive in the *pleistocene* still influence us, but they are not the whole story. Tooby and Cosmides would probably agree.

This sends a mixed message regarding the explanation of social norms. On the one hand, some norms are quite directly influenced by adapted modules. For instance, social norms regarding the treatment of feces (disgust, taboo, etc.) can probably be explained by a "brain circuit" that causes us to feel disgust when confronted with feces.

[6]Instead of repeating the arguments in the "sociobiology wars" (see Segerstrale, 2000; Kitcher, 1985), I move straight on to evolutionary psychology because this line of inquiry is, in my view, more fruitful for the discussion of norms.

This is a fitness-enhancing adaption in the EEA because contact with feces causes illnesses, and ultimately reduces replicative success. On the other hand, many other norms are probably better explained as products of culture and learning, and biological adaptation plays at best an indirect role in their explanation. For example, it would be surprising if conventions regarding the use of professional titles could be explained by referring to our genetical setup.[7]

5.3. Cultural Evolution and Norms

Cultural evolution might offer more direct explanations for social norms because cultural evolution moves much faster. This gives cultural evolution potentially more explanatory power with regard to the massive changes in social behaviour that occured with human civilisation.

Perhaps norms are "new replicators" in Dennett's sense: One can argue that they replicate through learning or imitation, they occasionally mutate, and they have differential fitness in the sense that some norms are more successful in replicating than others. However, there are two problems with this account. Firstly, the explanatory power of cultural evolution for social norms depends on whether there is a more substantive interpretation of fitness, rather than the tautological "whatever replicates faster is fitter". Secondly, it is unclear which "vehicle" norms use to replicate. One can either argue that norms latch on to single agents, influencing their behaviour. Or one can argue that norms are by their very nature universal, and their "vehicle" are groups rather than single agents. This determines the level of selection: Either norms are selected on the level of single agents, or on the level of groups.

I first turn to the definition of fitness. Henrich and McElreath (2003) give an excellent example of where fitness can be measured as success of the subjects using a norm: In 1860, a British expedition tried to cross Australia from north to south. The expedition ran out of food. Frequently, the travellers had to rely on presents from Aborigines, who were very successful foragers and had developed cooking techniques to process plants that are poisonous when unprocessed. Without the elaborate hunting and food processing techniques, which were passed on from generation to generation among the Aborigines, survival was impossible: "Humans, unlike other animals, are heavily reliant on social learning to acquire large and important portions of their behavioral repertoire."

[7] However, this does not mean that genetical adaptations play no role whatsoever for such norms. In the case at hand, issues of status management might be more directly influenced by adaptations because status was probably an important concept in the EEA.

(Henrich and McElreath, 2003, p. 123). The knowledge the Aborigines had made them successful survivors in the desert. The techniques they used were encoded in social norms ("this is how you hunt a fish", "this is how *nardoo* seeds must be processed"). In this example, the differential fitness of a norm is strongly positively correlated with the replicative success of the subjects complying with the norm. People who do not obey the right seed processing norm get sick and die, like most of the expedition party did before the Aborigines rescued the remaining survivors.

In many other cases, however, this link between the success of the subject obeying the norm and the replicative success of the norm is weak or questionable.[8] Clearly, humans obey many norms that do not maximise fitness for the norm compliers. For instance, in many African countries female genital mutilation is a common "cultural" practice.[9] Female genital mutilation does not enhance fitness in any objective sense. To the contrary, it endangers the health of its victims. So how is it possible that this practice has not been given up over time? The obvious answer is that social norms and cultural practices can prescribe all kinds of behaviour, including fitness reducing behaviour.[10] Researchers who are in favour of an evolutionary perspective on human behaviour often respond that cultural practices are not directly fitness maximising. The maximising mechanism behind the persistence of cultural practices is imitation. When an individual agent has to evaluate several courses of action, and if this evaluation is costly, it is often cheaper to copy other agents such as parents, authorities, peers, and so on (Boyd and Richerson, 1985, 2005). The systematic imitation of others is called *conformist transmission.* Conformist transmission is bad news if we want to predict which norms emerge by using an evolutionary analysis: If humans simply imitate others because this is epistemically beneficial, then the fitness of the norm they imitate is not a good predictor of its success. If imitation is the relevant mechanism, even highly inefficient and adaptively "unfit" norms can emerge and stabilise in a group.

In the case of female genital mutilation, we see a how a social norm is passed on from generation to generation because people tend to conform with existing social norms. This creates a lock-in effect: Cultural norms survive because "it has always been like that". This may be the result of an adaptation, namely the mechanism of conformist

[8] One obvious issue in that regard is the question of altruism: Truly altruistic behaviour reduces the fitness of the altruist and increases the fitness of the recipient. Can norms prescribing altruism survive in an evolutionary competition of norms? This leads to further debates about group selection or multi-level selection (see Sober and Wilson, 1998). I do not want to pursue these questions here. For now I merely want to expand the point that the differential fitness of a norm is often not linked to the fitness of subjects obeying it.

[9] See Mackie (1996) and `http://www.who.int/mediacentre/factsheets/fs241/en/`.

[10] For other examples, consider smoking or foot binding.

transmission. It is often efficient to copy others in certain environments, but this does not mean that the copied behaviour itself is fitness enhancing. Therefore one cannot conclude that cultural practices like female genital mutilation, which survive because of conformist transmission, are fitness enhancing for the people following them.

If this analysis is correct, then the link between fitness and norms is weak. This means that theories of cultural evolution have little or no predictive power for human behaviour driven by norms. To predict which norms are successful, or to explain why some norms have emerged while others have not, one needs a notion of fitness that differs from the replicative success of the norm. Otherwise the explanation reduces to the tautology: "it is successful because it is successful". For biological evolution we can explain and predict the adaptive fitness of a gene by analysing the phenotype and its interaction with the environment. But nothing similar can be done with the tautological fitness definition of Dennett's universal replicators or Dawkins's memes.

In the context of biological fitness, the tautology charge was rebutted by pointing out that fitness is not defined in terms of the actual replicative success, but rather in terms of expected reproductive success. This "propensity interpretation of fitness" (Mills and Beatty, 1979) avoids the tautology. In the context of cultural evolution, however, the tautology charge has more force. To use the propensity definition of fitness, we need to have a theory to estimate expected fitness. Such theories are available for fitness in biological evolution. But for fitness in cultural evolution, such theories are often lacking. If we want to predict the expected fitness of a social norm we run into the problem that almost any norm can succeed through conformist transmission. The distinction between actual and expected fitness avoids the tautology conceptually, but since we cannot estimate predicted fitness, we still end up with unsatisfying *ad hoc* and *ex post* explanations for why one norm was more successful than another.

Apart from the problem of defining fitness in a non-vacuous way, there is the problem of equilibrium selection. Even if there is a more substantial definition of fitness available, this does not necessarily mean that cultural evolution can predict which norms will emerge. For there are often many norms that would produce fitness advantages when obeyed, and it is unlikely that a model of cultural evolution could predict which of the alternatives should emerge as social norm.[11] In these cases, an evolutionary explanation can only show that the existing norm is fitness enhancing, but it cannot show why other possible norms, which are equally fitness enhancing, have not emerged instead.

I now turn to the challenge of finding the level of selection for the evolution of norms. In the standard picture of "memetics", "memes" are replicated by human imitation

[11] This problem is related to the problem of multiple equilibria in game theory. See Sugden (2001*a*).

(Blackmore, 1999). In Dawkin's metaphor this means that the "vehicle" for a "meme" are human agents. Consequently, we should expect a selection of "norm memes" on the level of individuals. But this is too simple. For an evolutionary analysis of norms we have to distinguish at least three different evolutionary processes: firstly, the process of *norm emergence*, starting with several norm candidates in competition within a society; secondly, the process of *norm enforcement*, where an existing norm is challenged by non-compliance; and thirdly, the process of *inter-group norm competition* between different norm systems.

These three processes entail three different evolutionary selection processes. I describe them in turn. The process of *norm emergence* takes place when a society needs a new norm for a new problem. At this stage, different candidate norms compete within the society. Different agents obey different norms. Some are more, some are less successful. Over time, the less successful agents adapt to the more successful ones, and a selection process occurs. *Norm enforcement* is a related but distinct process. When a norm emerges, it can face the challenge of non-compliance. Agents obeying the norm are in competition with those who do not obey the norm. If those obeying the norm do systematically worse than those who don't, then it is likely that the norm collapses over time. Therefore, the stability of a norm depends on how well it can be enforced. Finally, *inter-group norm competion* occurs when different groups follow different norms. The group with the fitter, more efficient norms has a selective advantage. Less successful groups might adapt the fitter norms or they decline if they are unable to adapt.

Once a norm is established and enforceable, it is coerced and almost universally obeyed. Norms are an example of near perfect horizontal information transmission. Everyone knows the norm, and everyone is expected to comply. Therefore, at this point selection is unlikely on the individual level: Even though there may be occasional mutations because people misunderstand a norm or violate it deliberately, there is no selection of "fitter" norms because any deviation from the prevailing norm is punished. One can argue that the norm produces its own "niche" by creating conditions which make obeying the norm optimal for an individual, independent from any objective standard of fitness. However, selection can still take place on the group level through the process of inter-group norm competition. A society does well when its norm system arbitrates conflicts efficiently, induces cooperation when needed, and coordinates behaviour such that the society benefits. In this sense, norm systems are under selection pressure on the group level.

The upshot is that the standard "meme" story is insufficient for norms because norms are not only selected on the level of single agents. Therefore, standard "memetics"

has no or only very limited use for analysing the evolution of norms. Instead, niche construction and group selection are the more appropriate frameworks. Also, since norms have a tendency to create their own niches, we should expect to find strong lock-in effects. Once a norm is in place it is hard to challenge it within a society. Change may happen through group selection, but group selection can be slow or ineffective when the fitness differences are not obvious.

These considerations should make us wary of overly ambitious explanations based on cultural evolution. Such explanations rest on feet of clay when they do not define fitness in a non-tautological sense. And even if there is a non-tautological definition of fitness, there are often several counter-factual but possible evolutionary stories that would be able to explain different outcomes. In addition, the theoretical insecurity about the level of selection and the questionable status of the "meme" metaphor and Dennett's "universal replicators" should induce caution. A group selection story for norms is theoretically more promising. But it still struggles with the fact that many norms do not appear to be efficient in any objective sense. For instance, female genital mutilation is spreading, rather than declining (Mackie, 1996). The norms prescribing it are not efficient in any objective sense, but they are frighteningly successful in creating a niche in which they thrive. Mackie calls this a "belief trap" (p. 1009). Since almost everyone in the relevant societies believes that female genital mutilation is necessary or even desirable, everyone supports it and no one can afford to question the norm. This shows that the explanatory leverage of cultural evolution is usually low when attempting to explain actual human behaviour. The upshot is that even if norms emerge evolutionarily, this does not suggest that they are necessarily objectively efficient or that evolutionary models can predict which norms emerge.

5.4. Thick Simulations: Evolution to the Rescue?

In chapter 4 I have argued that high fidelity models are usually not appropriate for the social sciences. Some computational modellers think that evolutionary theory can provide the information currently lacking for the construction of high fidelity models. In this section I revisit the issue of computational models in the light of evolutionary theory.

Timothy Kohler (2000) considers whether the knowledge derived from the theory of biological evolution can help us to move towards "thick" simulations (Kohler calls them "strong" simulations). Referring to Tooby and Cosmides (1992), Kohler observes that our "cognitive systems are seen as comprised of a large number of domain-specific

adaptations that were critical in the EEA [Environment of Evolutionary Adaptation], including things like face-recognition modules, and so forth" (p. 8). He then speculates that our increasing knowledge regarding evolutionary psychology should provide us with information on the dynamics of small-scale hunter-and-gatherer societies in their environment of evolutionary adaptation. However, when applying the same concepts to modern societies, problems arise: "outputs from these mechanisms, given ranges of inputs for which they were not evolved, become increasingly unpredictable" (p. 8). Therefore, modern societies, which are less driven by specific adaptations and more by social norms and cultural constructs, are more difficult to model. Nevertheless, Kohler concludes that

> "Despite these problems, it appears that an evolutionary psychological view of the world would open up the possibility of "strong social simulation" by focusing simulation efforts on *evolving* adaptations and mechanisms for adjudicating among adaptations." (Kohler, 2000, p. 8)

This careful optimism is shared by Herbert Gintis (2007) in an effort to present "a framework for the unification of the behavioral sciences". Gintis aims to link evolutionary theory (including theories of co-evolution) with decision theory. He builds a bridge between the two by arguing that consistent choices lead (on average) to higher adaptive fitness (p. 4). Gintis is at pains to balance this claim properly: On the one hand, his "unification" project needs a link between evolution and decision theory, and this link is provided by "fitness":

> "Brains, therefore, are ineluctably structured to make, on balance, fitness-enhancing decisions in the face of the various constellations of sensory inputs their bearers commonly experience." (Gintis, 2007, p. 3)

On the other hand, Gintis makes an effort to emphasise that humans are not always maximising their fitness, neither consciously, nor unconsciously:

> "I have argued that we can expect the BPC [beliefs, preferences and constraints model] to hold because, on an evolutionary time scale, brain characteristics will be selected according to their capacity to contribute to the fitness of their bearers. But, fitness cannot be equated with well-being in any creature. Humans, in particular, live in an environment so dramatically different from that in which our preferences evolved that it seems to be miraculous that we are as capable as we are of achieving high levels of individual well-being." (Gintis, 2007, p. 10)

The problem is that Gintis does not say which definition of fitness he wants to use. In the first quotation he suggests a notion of fitness in the biological sense. If this is his preferred notion of fitness, Gintis has to explain how to make sense of all the human behaviour that is not fitness maximising. As argued above, many social norms are not fitness maximising, but persist nevertheless. However, if he works with a weaker notion of fitness, as he seems to suggest in the second quotation, then such a theory does not have much explanatory power. I have argued above that the fitness of a norm is often not linked with the fitness of the subjects applying it. This can clearly be seen by looking at cases such as female genital mutilation, smoking, etc. Gintis's claim that brains are "ineluctably structured to make, on balance, fitness-enhancing decisions" is therefore questionable. And even if there is a link between the fitness of the norm complier and the replicative success of the norm, this does not mean that it provides a working explanatory theory because there might be many conceivable fitness enhancing norms, and the theory would provide no criterion for distinguishing between them. In any case, without a clear definition of fitness one cannot evaluate Gintis's "unification project".

Returning to the case of the Anasazi, the help we can expect from evolutionary psychology to derive "thicker" simulations is limited. The fate of the Anasazi population in the Long House Valley was probably to a large extent determined by their specific cultural practices and social norms, and the interaction between norms and natural environment. While evolutionary psychology may provide us with some ideas as to how the Anasazi behaved, we are unlikely ever to reconstruct how their social norms shaped their specific way of life and how this in turn influenced their growth and rapid decline. When social processes are influenced by rich cultural phenomena, it is unlikely that we are able to devise high fidelity simulations of these processes.[12] Upon closer inspection, this is not surprising. We know, after all, that all kinds of norms survive, and that many of them are patently stupid or dangerous. To return to the example of female genital mutilation: The norm prescribing this outrageous procedure obviously have high differential fitness in some societies. But this does not imply that applying this norm conveys any fitness to its practitioners or its victims.

At the bottom line, cultural evolution is a useful framework to think about the evolution of ideas and behavioural patterns. This may lead to useful *ex post* explanations. For instance, Dennett's analysis of religions is a case in point (Dennett, 2006). But

[12]Ironically, evolutionary psychology might be of greater help for the Redfish stadium model. The instincts humans apply in crisis situations are probably to a certain extent genetically hardwired and evolutionary psychology could provide ideas about how humans might behave.

cultural evolution is not a good framework for *predicting* which "memes" will survive or emerge. Consequently, one should take claims that an evolutionary perspective helps to build high fidelity models with a large pinch of salt.

5.5. The Evolution of Norms: Grand Theory or Small Games?

It is of great use for this discussion that Ken Binmore has published "Natural Justice", a forceful, provocative, and sharply argued theory of evolutionary justice (Binmore, 2005). Binmore leaves no doubt about what he thinks about moral philosophy in general, and Kantian philosophy in particular:

> "We are told that we must adopt the prejudices of famous philosophers like Kant, Moore, or Rousseau because of the demands of Practical Reason, Moral Intuition, or the General Will. A metaphysical Nature, unrelated to the natural world of plants and animals, is sometimes wheeled on to provide a justification for our having Natural Rights, and being subject to Natural Law." (Binmore, 2005, p. 44)

In an earlier interview he defends his evolutionary naturalism against the critics:

> "I am a total thoroughgoing naturalist, reductionist, relativist, all those naughty things. Asking 'how ought we to live?' is like asking 'what animals ought there to be?' Like the animals that exist, the moral rules that we have are shaped largely by social and biological evolutionary forces. If one wishes to study morality it therefore makes no sense to ask how moral rules advance the Good or preserve the Right. One must ask instead how and why they survive." (Voorhoeve, 2002)

This is the most radical naturalistic and evolutionary approach to morality and norms I am aware of. Binmore ridicules the idea that there might be any non-natural moral properties. He also claims that all moral properties can be reduced to natural properties. In fact, his theory is even stronger. He rejects the idea that the question of how we *ought* to live makes any sense, apart from reducing it to propositions about how humans evolved in the natural process of evolution. Therefore, according to Binmore, what is good and right is determined ultimately by what has survived in evolution. However, Binmore does not deny that there is justice in the world—but his justice is a "Natural Justice", the title of his latest book.

5. *Some Remarks on Evolutionary Explanations of Norms*

Binmore takes no prisoners, attacking "rationalists" and their "Humpty Dumpty" philosophy. It is easy to be provoked by Binmore's polemics, his lashing out against Kantian "Alice-in-Wonderland reasoning" (Binmore, 2005), his unwillingness to engage with arguments he does not like. However, leaving polemics aside, Binmore has something important to say. In effect, he proposes to invent moral and political philosophy again, this time on naturalistic grounds. This is a paradigmatic example of what I call "grand sociobiological theory". Sugden (2001*b*), reviewing Binmore's earlier voluminous "Game Theory and the Social Contract" (Binmore, 1994, 1998), criticises the whole project. He emphasises that Binmore is engaged in a major attempt to "recover" moral and political philosophy on naturalistic grounds. This naturalistic "recovery project" tries to rebuild ethical theory without making any *a priori* claims or normative assumptions. Binmore dismisses all attempts to ground ethics on rationalism, natural law and similar concepts as "skyhooks" without justification. Instead, he argues, the justifications for ethical conduct should be inferred only from natural (and in particular evolutionary) processes. If the naturalistic "recovery project" succeeded, moral philosophy would be transformed into a natural science like chemistry or biology.

Sugden's critique is devastating, arguing that Binmore's ambitious "recovery project" fails on several accounts. Put in a nutshell, Sugden doubts that Binmore succeeds in recovering concepts such as "rationality", "bargaining", "Nash equilibrium" or "fairness" on naturalistic grounds. He also criticises Binmore's unwillingness to engage with empirical research. Clearly, a naturalistic reconstruction needs empirical support, since *a priori* reasons are not available.

While Binmore sees his theory as the spearhead in an all-out attack against moral and political philosophy, other authors tread much more carefully. Brian Skyrms (1996; 2004) starts his exploration of evolutionary processes with iterated two-person games. By combining evolutionary replicator dynamics with simple game-theoretical models, Skyrms explores the elementary mechanisms of how evolution could have produced norms of cooperation and equality.

Skyrms pleads for a conciliatory view regarding the relation between moral philosophy and evolutionary research:

> "Ethics is a study of possibilities of how one might live. Political philosophy is the study of how societies might be organized. If possibility is construed generously we have utopian theory. Those who would deal with 'men as they are' need to work with a more restrictive sense of possibility. Concern with the interactive dynamics of biological evolution, cultural evo-

> lution, and learning provides some interesting constraints." (Skyrms, 1996, p. 109)

Skyrms does not offer an overarching theory as to how evolutionary game theory can inform moral philosophy. The quote is perhaps his most explicit metaethical statement in "The Evolution of the Social Contract". His distinction between "utopian theory" on the one hand, and dealing with "men as they are" on the other, allows Skyrms to assign a special role to evolutionary game theory: Game theory, used to explore the "interactive dynamics of biological evolution, cultural evolution, and learning" reveals the "constraints" under which realistic moral philosophy has to operate. Utopian theory is unconstrained in its prescriptions. Realistic theory can only prescribe what is practically feasible for moral agents (see chapter 3).

Skyrms's constraints are evolutionary constraints in a wide sense. He believes that evolutionary dynamics drive biological evolution, cultural evolution, and learning. Biological evolution implies indirect constraints: Human physiology and hardwired aspects of human psychology restrict realistic moral philosophy. For instance, if we assume that humans have a genetically determined aggressive drive, then realistic moral philosophy should take this into account and expect norms to contain aggression.[13] More important sources of constraints, however, are cultural evolution and learning. Again, Skyrms is not very explicit about the link between cultural evolution and realistic moral psychology. Skyrms, well aware that literally all aspects of cultural evolution and its influence on moral theory are hotly contested, wisely avoids too much exposition.

Roughly speaking, Skyrms seems to argue that the dynamics of learning and cultural evolution constrain realistic moral theory because evolution and learning tend to promote a spread of more "successful" behavioural patterns. His considerations operate on the level of norm emergence, not norm enforcement or inter-group norm competition. This is how I understand the mechanism that Skyrms might have in mind: A group of agents has to interact with each other over a longer period of time. Agents can either follow behavioural pattern A or behavioural pattern B (A and B are strategies in game-theoretical terms). Those who follow behavioural pattern A are more "successful" in their interactions, compared to those who follow B. "Successful" can mean that they obtain greater material resources (food, money, etc.), but any form of "success" would do for the argument, as long as all agents care about being successful. Now enter the dynamics. When agents are able to learn, they can change their behavioural patterns over time. In the example, those who start with behavioural pattern B will

[13] This is my example, not Skyrms's.

observe that those with pattern A are more successful. Therefore, they change their behaviour and adapt pattern A, too. Learning dynamics increase the frequency of the more "successful" behavioural pattern over time.

Different processes can drive the dynamics, as discussed above. Imitation and vertical information transmission probably play a role. Niche construction may also be important, especially for social norms of cooperation, as seen above. The details of the dynamics are not crucial for the core argument and Skyrms is reluctant to discuss details. What matters is that learning dynamics and cultural evolution can work as a selection mechanism for behavioural patterns. How does this constrain realistic moral theory? The reasoning works as follows: Firstly, if the agents desire "success" (and agree on a definition of "success"), and if they can change their behavioural patterns over time, then behavioural patterns that do systematically worse will disappear. Secondly, social and moral norms can cause behavioural patterns. If Norm N_A prescribes actions according to behavioural pattern A, and norm N_B prescribes actions according to behavioural pattern B, then the competition between the two behavioural patterns turns into a competition between two different norms. In a population where some agents obey N_A and others N_B, the agents following N_B systematically "lose out". Agents can adapt over time: They change their norm system and give up N_B in favour of N_A. Consequently, norm N_A prevails and N_B "dies out". A realistic moral theory cannot prescribe N_B because it does not survive the evolutionary competition with norms like N_A.

This simple account of evolutionary norm emergence needs several qualifications. Firstly, the dynamics can be more complex. More norms compete with each other, and they might differ in the complexity and conditionality of behaviours they prescribe. Secondly, which norms compete with each other is very difficult to predict *ex ante*. Thirdly, when and how moral agents update their moral beliefs is primarily an empirical question, and the process is likely to be much more difficult in reality. This list could be continued. Nevertheless, despite the simplicity and its obvious shortcomings, a model of evolutionary norm competition is not useless. I think that Skyrms is right in emphasising the importance of constraints. However, in contrast to Skyrms, I propose to shift the debate away from the question of norm emergence and focus on the issue of norm enforcement. What we should study are the sanctioning mechanisms that enabled currently existing norms to do well. We are usually not able to observe the genesis of new norms—what we are able to observe is how existing norms are enforced. Understanding sanctioning and enforcement mechanisms enables us to explain how existing norms have survived evolutionary competition.

5.6. Evolutionary Constraints

I have argued that the predictive power of evolutionary theories regarding norms is limited at best. But this does not mean that evolutionary theories are useless for the analysis of social norms. What we have to do is change the question: Rather than asking for predictions, we can more productively ask why and how our present social norms have survived, given that they experienced competition from competing norms and from the challenge of non-compliance. For *social norms of cooperation*, one particularly interesting question is how they managed to stand their ground when there is a constant threat of non-cooperative defection. Somehow these norms must rely on sanctioning mechanisms. To analyse these mechanisms, it is not always necessary to model the evolutionary process itself. What is primarily of interest is the product of evolution, not the process. In chapters 6 and 7 I will analyse variants of social exclusion as a subtle enforcement mechanism, but I do not model an evolutionary process directly. In other words: The research question for my models is based on evolutionary considerations, but my models are not evolutionary.

Evolutionary theories are often used in more ambitious projects. Binmore's naturalistic ethics may be just the tip of the iceberg.[14] I am skeptical that such "grand sociobiological theories" stand up to scrutiny. In my opinion, Skyrms's small models and his circumspect and careful way to draw inferences about the real world are sounder. Small models are advantageous when the target system is not well understood (see chapter 4). In addition, I have shown above that the explanatory power of evolutionary theory is hampered by its limited predictive power: If there is no single conception of fitness, or if the definition of fitness is tautological, then there is little predictive power in an evolutionary theory of norms.

In chapter 3 I have claimed that dynamic models of norm compliance usually rest on the assumption that people are at least partly driven by self-regarding utility maximisation and switch to selfish strategies when they think that many others are selfish. Evolutionary theory is a possible justification for these assumptions because evolutionary pressure should lead to utility maximising agents, and (at least in infinite encounters) conditional cooperation is a sensible strategy. This argument works well when people's utility is closely linked to their material well-being and fitness depends on material well-being. In this situation, an evolutionary selection process produces norms with high fitness, which in turn will be norms that secure high levels of material well-being,

[14] A different, in my opinion more carefully argued but still staunchly naturalistic example is Michael Ruse's (1995) work.

which in turn produces high utility for the agents, which in turn produces high fitness for the agent. Here, the fitness of the norm, material gain, agent utility, and the fitness of the agent are all in line. In these situations, evolutionary reasoning works best.

However, consider a case where agents derive utility not from material gains, but from the esteem they receive from other people (see Brennan and Pettit, 2004). Assume that esteem depends on how well agents comply with relevant social norms. People are still maximising their utility (and are rational in that regard), but now the fitness of the norm comes apart from the fitness of the subject applying the norm. Obeying the norm increases the subject's utility, but it does not increase its objective fitness. For instance, there might be a norm that demands sacrificing 10 percent of one's harvest to the gods. Complying with this norm triggers esteem, non-compliance disesteem. It is perfectly rational for people in this society to sacrifice food, even though this diminishes their material goods and their objective fitness. Such norms are often stable, and the *norm* has a high replicative success. But an evolutionary approach would not have predicted the emergence of the sacrifice norm because the fitness of the *agent* is sub-optimal.

Where does this analysis leave us regarding evolutionary models of norms? Jason Alexander subscribes to a skeptical view:

> "If one wishes to explain how some currently existing social phenomenon came to be, it is unclear why approaching it from the point of view of evolutionary game theory would be particularly illuminating. The etiology of any phenomenon is a unique historical event and, as such, can only be discovered empirically, relying on the work of sociologists, anthropologists, archaeologists, and the like. Although an evolutionary game theoretic model may exclude certain historical sequences as possible histories (since one may be able to show that the cultural evolutionary dynamics preclude one sequence from generating the phenomenon in question), it seems unlikely that an evolutionary game theoretic model would indicate a unique historical sequence [that] suffices to bring about the phenomenon. An empirical inquiry would then still need to be conducted to rule out the extraneous historical sequences admitted by the model, which raises the question of what, if anything, was gained by the construction of an evolutionary game theoretic model in the intermediate stage." (Alexander, 2003*a*, section 4.2)

I concur with Alexander that the *exclusion* of historical sequences is currently the best we can do with evolutionary models of social behaviour. Exclusion does not give us

explanations of social behaviour as it is, but it may give us at least the feasibility constraints for social behaviour. I have discussed the importance of feasibility constraints for a descriptive theory of morality in chapter 3. Insofar as evolutionary theory can offer an analysis of constraints in Alexander's and Skyrms's sense, it has a useful function for understanding our actual moral systems and social norms. I have argued above that even this limited role comes with difficulties, because the explanatory power of cultural evolution suffers from ambivalent definitions of fitness. Still, evolutionary theories of morality and social norms at least help us to ask the right question: If norms prescribe actions that reduce the objective fitness of the agent, how did these norms survive? To answer this question and understand social norms, we must look at the subtle mechanisms of norm enforcement.

5.7. Conclusion

This chapter started with a short outline of biological and cultural evolution. I then moved on to discuss the importance of both theories for the analysis of social norms. Cultural evolution has a more direct relevance for social norms because most social norms are—in the time frame of biological evolution—too young to explain their development and changes biologically. Norms may be "new replicators" in Dennett's sense. They replicate, mutate, and have differential fitness. However, the measurement of differential fitness is one of the core controversies in the field of cultural evolution. If the fitness of a norm is nothing but its ability to replicate in comparison with competing norms, then the notion of fitness is vacuous, and the explanatory power of evolutionary theories of norms is low. I argue that more substantive interpretations of fitness run into problems because many norms are not fitness maximising in this deeper sense. For this reason, I also argue that evolutionary psychology should not be overestimated in its ability to support high fidelity models of social behaviour. In addition, there are at least three different processes related to the evolution of norms: the emergence of norms, the enforcement of norms, and inter-group norm competition. The simple "meme" story alone does not work.

Binmore's evolutionary theory of justice is a good example to see the dangers of overstretching evolutionary theories to explain human behaviour. Less ambitious, but—I think—ultimately more fruitful is Skyrms's approach. Small, highly idealised evolutionary models can be used to explore the constraints for moral systems and other norms. If norms demand too much, they become unrealistic and cannot be enforced, given that they face competition from other, less demanding norms and the challenge

of non-compliance. This approach changes the question: It does not try to explain why specific norms have emerged, it rather asks how existing norms have managed to survive in evolutionary competition with other norms. One important part of the answer are sanctioning mechanisms: Norms survive if they can be enforced. In chapters 6 and 7 I investigate exclusion and group formation as subtle mechanisms of norm enforcement.

Part II.

Formal Computational Models of Norm Enforcement

6. Assortation and Translucency[1]

6.1. Introduction

In part I of this thesis I have developed the notion of realistic moral theory, arguing that feasibility of moral actions should play a role in moral theory. More specifically, I argued that social norms of cooperation face a dilemma regarding compliance, and that it is important to understand the subtle mechanisms of norm enforcement. In what follows I analyse one such enforcement mechanism: social exclusion. It turns out that social exclusion can be a subtle, yet highly effective sanctioning mechanism. It is likely that the dilemma of norm compliance is often solved by threatening the exclusion of norm violators.

This chapter discusses mechanisms that secure high levels of cooperation in public goods dilemmas with anonymous contributors. Since anonymous public goods dilemmas have payoff structures in the form of n-person one-shot prisoner's dilemmas, cooperation seems unlikely. With payoff structures of this form, cooperators fare worse than defectors. However, Skyrms notes that cooperative strategies perform better if strategies are positively correlated in their encounters, i. e. if cooperators are more likely to meet other cooperators (Skyrms, 1996, p. 45–62). To achieve positive correlation, two conditions must be met. Firstly, agents must be able to recognise the strategies others play. Secondly, they must be able to select the other agents they play with.

The first condition is met when agents are translucent (Gauthier, 1986, p. 174–177; see also Frank, 1987; Frank, 1988; Sally, 2000). Translucency means that agents can fallibly predict which choices other agents make in unrepeated non-cooperative games, such as the prisoner's dilemma. The second condition is met by a mechanism of group formation. Sethi and Somanathan call this mechanism "assortation" (Sethi and Somanathan, 2003). There are a number of articles using translucency and assortation to explain cooperation (Güth and Kliemt, 2000; Congleton and Vanberg, 2001; Page, Putterman and Unel, 2005). The problem with many of these approaches is that they

[1] An earlier version of this chapter was published in Politics, Philosophy & Economics (Spiekermann, 2007).

either have to assume unrealistically high levels of translucency, or they have to assume repeated interactions with learning effects, failing to address the challenge of the most severe dilemmas. This chapter takes a different route: Agents face a real one-shot dilemma, ruling out learning effects. Instead of assuming unrealistically high levels of translucency, the chapter shows that even weak translucency suffices when groups of agents pool their information. Using the insights from the Condorcet Jury Theorem, I argue that weak translucency and assortation secure cooperation when assortation decisions are based on collective rather than individual judgements. This underscores the importance of groups and collective judgements to reach high levels of cooperation under adverse conditions.

The argument has close links to David Gauthier's "Morals by Agreement". In a related argument, Robert Frank has argued that emotions could play an important role for translucency because they are difficult to feign (Frank, 1988, ch. 5). In addition, there is a widespread literature on the possibility of altruism, which is linked to the question of cooperation under adverse conditions. Evolutionary philosophers in particular try to determine whether "hard-core altruism" (E. O. Wilson cited in Frank, 1988, p. 25) can survive in an evolutionary setting. However, none of these earlier contributions takes the epistemic advantages of groups into account. This chapter shows that it makes sense for cooperators to care about the cooperative disposition of other people.[2] The chapter is in seven sections. I begin by discussing the problems faced when modelling moral dilemmas in a rational choice framework (6.2). I then introduce translucency and assortation as the relevant concepts to protect cooperators from free-riding (6.3). I begin with a simple model to explain the general idea (6.4) and move on to a more complex model (6.5), which is then put to use in a computer simulation (6.6). I discuss results and remaining challenges in section 6.7 and conclude in section 6.8.

6.2. The Costs of Cooperation

The question under which conditions cooperation can emerge in adverse environments has received intensive attention from game theorists, political theorists, and philosophers. Table 6.1 shows a one-shot prisoner's dilemma (PD). More precisely, the table

[2]The interest in cooperative intentions and the need to communicate about it also connects translucency with the phenomenon of esteem. See Brennan and Pettit, 2000, 2004. Holding someone in esteem is a signal that the "esteemer" has acknowledged the integrity of the esteem target (and is interested in future cooperation). The desire to be held in esteem is the desire that others observe and acknowledge one's integrity (and are consequently willing to interact in future games).

	cooperate	defect
cooperate	2, 2	0, 3
defect	3, 0	1, 1

Table 6.1.: A prisoner's dilemma.

shows a PD if the numbers in the table represent preferences over outcomes. If this is the case, then table 6.1 defines a game by stating the players, the strategies available to players, the outcomes, and the players's preferences over the outcomes. Both prefer the outcome "I defect, the other cooperates" over "both cooperate" over "both defect" over "I cooperate, the other defects". Thus, defection strictly dominates cooperation for both players, and mutual defection is the only Nash equilibrium.

Game theorists emphasise that the dynamics of 2-person PDs change fundamentally when they are repeated indefinitely. Robert Axelrod's computer tournaments have spurred the debate. Axelrod (1984) argues that cooperative strategies can succeed in indefinitely repeated games. In particular, Axelrod shows that TIT-FOR-TAT is a simple but very successful strategy to establish lasting cooperation in repeated 2-person PDs. It is questionable whether TIT-FOR-TAT deserves the special attention it has received (Binmore, 1994, p. 194; see also Nowak, 2006*a*, p. 78–91), but it is clear that cooperative agents can prosper in repeated dilemmas, while the corresponding one-shot games prove disastrous for cooperators. In fact, any pure or mixed pair of strategies played in an indefinitely repeated PD can be in a Nash equilibrium according to the Folk theorem.[3] The implication is that classical game theory can explain cooperation in indefinitely repeated games. However, if a situation is truly a one-shot PD, defection is the analytically rational strategy for both players, because a PD, expressed in terms of all-things considered, immediately action-guiding preferences, logically implies that both players defect.

Unfortunately, the fact that PDs imply that both agents prefer to defect has often led to the false belief that individuals always defect when they face a game form with a *monetary or material payoff structure in the form of a PD.* If the above matrix shows all-things-considered preferences, then rational agents defect. But if the above matrix shows monetary payoffs, it remains an open question what agents do (and the matrix would only state a game form, not a game, as the players's preferences are not yet defined). The monetary payoff structure does not determine the agents' preference

[3]See Gintis, 2000*a*. A Folk theorem with subgame perfection even holds for repeated n-person games. "However, the existence of vast numbers of subgame perfect Nash equilibria renders the repeated game framework virtually useless" in Gintis's (p. 129) opinion.

structure. In this chapter I explore the behaviour of agents who face material payoff structures in the form of a PD, but have different preferences of how to act in this situation. I discuss two types of agents, *cooperators* and *defectors*. Cooperators, when challenged with the material payoff structure as above, prefer to cooperate when the other player cooperates, but they prefer to defect when the other player defects. They play a coordination game. Defectors, by contrast, play a PD and always defect.

Cooperators are the more interesting agents. Rather than maximising their own monetary payoff, they try to realise coordination gains. I assume that cooperators feel committed to a rule of cooperation. However, such rules are tested against reality. If it turned out that cooperators fared badly in material terms (compared to defectors), the rule to cooperate could hardly be sustained. People can only afford to be "nice" when their self-interest does not suffer too much. Thus, cooperators try to be cooperative, but they still face the material payoff structure in the form of a PD. To turn cooperation into a success story, cooperators must protect themselves from exploitation, otherwise they are forced to abandon the rule of cooperation.

Agents can follow a rule of cooperation because of different motives. In this chapter I focus on a moral commitment to cooperation. I also use the term "integrity" to express that agents are motivated by moral obligations to cooperate. However, the problem I discuss can be extended to the more general problem of social norm compliance: The function of social norms is to prescribe actions that would not necessarily be chosen if agents acted only in their own interest (Bicchieri, 2006, p. 11). If compliant agents interact with non-compliant ones, the compliant agents are disadvantaged, because they restrain their pursuit of self-interest, while the non-complying ones don't. Another background theory to motivate a rule of cooperation is McClennen's "resolute choice" (1990). The idea would be that agents have chosen to cooperate and resolutely follow this plan despite the temptation to defect.

This chapter shows that cooperators can protect themselves and play cooperatively, despite the threat posed by free-riding defectors. The environment I use to test the viability of cooperation is hostile to cooperation: Agents face a payoff structure in the form of an anonymous n-person prisoner's dilemma (ANPD). More precisely, they play a public goods game where individual contributions cannot be observed. This offers a great opportunity to free-ride because later punishment is impossible, given the anonymous nature of the game. However, the situation of cooperators is not hopeless when we give them the chance to shape the environment of interaction, using translucency and assortation. These two mechanisms, taken together, provide cooperators with the means to protect themselves against free-riding.

6. *Assortation and Translucency*

While game theorists have explored the mechanisms leading to the possibility of sustained cooperation in repeated PDs, others have pointed out that these results are of limited use for *moral* theory. Game theorists show that cooperation can be in an agent's self-interest if the game is repeated indefinitely. While these results are important, they cut no ice regarding morality. There is a difference between our self-regarding interest (the interest in our personal welfare), and the requirements of morality. This difference can be seen by considering the reasons for our actions. It is suitable to justify non-moral actions by claiming that they promote our own welfare. However, to justify moral actions, we need to transcend this purely self-regarding perspective, and refer to other-regarding or impartial reasons. Even though it is possible that our self-interest coincides with the requirements of morality, self-interest is not an adequate reason for moral action. Moral reasons override reasons derived from self-interest in the case of moral action: The point of morality is that it commits us to actions that may go against our self-interest. Morality often forces us to resist our self-interested desires in favour of our moral commitments. If moral action was always self-interested, we would not be able to understand what a moral dilemma feels like. Danielson concludes:

> "[Tit-for-tat] is not a moral principle, because in the Iterated Prisoner's Dilemma it is straightforwardly in an agent's interest. [...] Since straightforward maximization suffices here, there is no need for a new kind of principle, namely a moral principle constraining an agent's self-interest." (Danielson, 1992, p. 46)

Is there a better way to integrate moral commitment into rational choice theories? On the one hand, it is unsatisfying to model moral commitment as a preference to act morally. In an important sense we do not prefer to act morally, but rather feel compelled to act morally despite our more selfish desires (Sen, 1977).[4] On the other hand, decision theorists are likely to argue that our ultimate preferences, which determine our decisions, must incorporate our moral considerations. Our desire to act morally might contradict our desire to further our self-interest, but if we decide in favour of morality we *prefer* the moral action over the self-interested action (see also McClennen and Shapiro, 1998).

The question is how to model the challenge of moral dilemmas, without throwing standard assumptions of rational choice theory out of the window. Philip Pettit (2000)

[4] My concept of "commitment" is similar to Sen's. A commitment "drives a wedge between personal choice and personal welfare" (p. 329). By contrast, acting from "sympathy" always maximises one's own (expected) welfare.

suggests interpreting rational choice theory not as a general theory for every human behaviour, but rather as a theory of constraints, such that deviations from the rational course of actions are corrected, but only if these deviations lead to substantial losses in self-regarding terms. He suggests that rational choice theory assumes primarily self-regarding preferences, while actual human behaviour is often not driven by self-regard. However, self-regard is important in a virtual way:

> "Under the model of virtual self-regard, most actions are performed without self-regarding consideration but that is true only so far as most actions happen to do suitably well in self-regard terms. The agent is genuinely moved by ordinary, culturally framed considerations but only so far as those considerations do not require a certain level of self-sacrifice. Let the considerations push the agent below the relevant self-regarding level of aspiration—this will vary, no doubt, from individual to individual—and the alarm bells will ring, causing the agent to rethink and probably reshape the project on hand." (Pettit, 2000, p. 41)

Pettit's model eases the pressure on rational choice theories to explain every instance of human behaviour. According to Pettit, rational choice theory provides us only with a baseline as to how individuals behave approximately. They may exhibit all kinds of action not predicted by the theory, but when substantial material interests are threatened, self-protection becomes the dominant desire, which is well-predicted by rational choice theory. Rational choice theory does not predict every behaviour, it merely limits the set of possible behaviours.

Pettit's theory of virtual self-regard provides a blueprint for the analysis of moral behaviour in terms of rational choice theory. On the one hand, agents face a material payoff structure, implying a strategy for individual payoff maximisation. On the other hand, agents have moral commitments, which may go against the maximisation of material payoff. Agents can afford to stick to their commitments, as long as their material well-being is not systematically at risk. If a norm demands actions leading to systematic disadvantages, it does not pass the realism test and will be abandoned.

6.3. Translucency and Assortation

Two processes help cooperators to keep defectors in check: Translucency and assortation. I discuss both in turn.

Translucency

When the payoff structures (not the preferences) take the form of a one-shot PD, it seems that defection is the only viable strategy to play, if the success or survival of strategies depends on their material success. Brian Skyrms refutes this claim. Skyrms (1996, p. 45–62) shows that cooperative strategies stand a chance in such settings where the model allows for positive correlation in the encounters of strategies. This means that cooperative strategies can survive if they have a better than random chance to play against other cooperators. Under suitable conditions, cooperators manage to interact only with their own kind, while defectors are forced to interact only with defectors. In that case, cooperators achieve higher payoffs because they reap coordination gains inaccessible to defectors.

If correlation is perfect, cooperators only meet cooperators, defectors only defectors. With perfect correlation, cooperators do much better than defectors. Even with imperfect correlation, cooperators can have an advantage. The question is then: Can positive correlation emerge in human interactions? Some authors think it can. Frank (1988, ch. 5) assumes that emotions reveal agents' dispositions to cooperate or defect. Gauthier (1986, p. 174–177) uses the concept of "translucency" in his "Morals by Agreement", claiming that individuals can fallibly distinguish those who are committed to cooperation from those who are not. Sally (2000) describes a similar phenomenon, which he calls "mind-reading".

Some authors have conducted empirical tests of translucency. The results support the assumption that agents can detect the cooperative disposition of others even when the *ex ante* interaction is short. Frank, Gilovich and Regan (1993) show that individuals improve their prediction whether their opponents cooperate, when they have a 30 minute discussion before the game. This holds even though the experiment prevents participants from discovering who has defected *ex post*, i. e. when punishment is impossible and defection is the strictly dominant strategy. The results imply that a) discussion produces or activates rules of cooperation; b) many agents stick to these rules; and c) individuals improve their prediction of whether their opponents stick to rules of cooperation, when they have the opportunity to meet them before the game. Brosig (2002) obtains similar results, while Ockenfels and Selten (2000) are sceptical.

Assortation

Translucency alone is not sufficient to generate positive correlation. To increase the chances that cooperators meet other cooperators, some social framework is necessary.[5] Cooperators must be able to shape their environment such that the information gained through translucency increases their encounters with other cooperators. One possible mechanism is a process of assortation (Sethi and Somanathan, 2003). Before agents play, they form groups. In this group formation process, cooperators can influence which other agents they will play with. Given translucency, cooperators have the opportunity to choose other cooperators. Therefore, positive correlation is created by forming at least two groups: An "in-group" with a high proportion of cooperators, and an "out-group" with a low proportion of cooperators.

If translucency is strong enough, and if the social mechanisms are set up in a suitable way, the level of positive correlation might be high enough for cooperators to do well in material terms.

6.4. A Simple Model

To explain how assortation and weak translucency can help to maintain cooperation in hostile environments, I start with an informal description of the simplest setup. Let there be an infinite pool of (potential) agents. Out of this pool a group of n agents emerges to play a game with an anonymous n-person prisoner's dilemma (ANPD) payoff structure. More precisely, let the ANPD be a public goods problem. Each agent can either contribute one or zero units to a group fund. The fund is multiplied by k, with $1 < k < n$ and is then divided equally between the n members of the group.

There are two types of agents: *cooperators* and *defectors*. Defectors understand that mutual defection is the Nash equilibrium in terms of payoff. They defect no matter what other agents do. Cooperators have a "nicer" disposition: They are willing to cooperate if they believe that their payoff in the game is positive. Cooperators play a coordination game. They cooperate if there are "enough" other cooperating agents. "Enough" means that the ratio of cooperation in the group is greater than $1/k$, which yields a positive payoff.[6]

[5] I concur with Brennan (1996) who observes that economic theory has underestimated the incentives in *ex ante* selection processes compared to *ex post* incentives.

[6] This behaviour is often called "conditional cooperation" in the economic literature. There is now reliable evidence that many people use strategies of conditional cooperation when they face public goods dilemmas. For a review of recent experiments and further references see Gächter (2006).

6. Assortation and Translucency

It is important that cooperators do not ignore the monetary payoff structure determined by the game form. The game form, which defines the payoffs for different outcomes, is the same for cooperators and defectors. The games they play, however, are different because the two types have different preferences over outcomes. Defectors play a PD and always defect. Cooperators play a coordination game because they do not only consider the situation from a payoff-maximising, but also from a rule-guided perspective. Cooperators believe that they are under an obligation to cooperate with other cooperators, and that defection against other cooperators is wrong. Having said that, cooperators cannot afford to be constantly exploited and make monetary losses. Thus, cooperators need to make sure that they do well enough on the monetary side.

The question is how cooperators can safeguard their material well-being if they play an ANPD in terms of the payoff structure. Here is the idea: Assume that every agent has a weak ability to recognise the disposition (cooperator or defector) of every other agent. This means that each agent receives a signal on the disposition of every other agent, which is better than random, i. e. the likelihood that the signal is correct is greater than 0.5, conditional on the true disposition of the observed agent.

As the competence to recognise the disposition of other agents is only slightly better than random, there is only weak translucency. The information quality for each single agent is poor. This is bad for cooperators because it lowers their ability to determine whether they can cooperate without experiencing a monetary loss. Unless they can be reasonably sure to play with enough other cooperators, they must defect to avoid exploitation.

However, the situation of cooperators improves if they can pool their individual signals and use this pooled information to avoid interaction with defectors. Assume that agents decide by simple majority voting which agents to accept as members of their group. Both cooperators and defectors want to have more cooperators in the group. Assume that each additional member is accepted or rejected by a vote, and that all agents vote informatively, that is they reveal their private signals truthfully. Each member of the group is competent, i. e. the signals received are better than random. Moreover, strategic voting is ruled out, and the judgements of group members are independent. In this setting, the Condorcet Jury Theorem applies.[7] Each decision to accept a new group member is a jury decision in Condorcet's sense. The existing group members judge the candidate. If their private signal indicates a cooperative disposition, they vote in favour of the candidate, if not they vote against. As the group size grows, the collective competence of the group to identify cooperators approaches 1. This in

[7] For a discussion of the Condorcet Jury Theorem see List and Goodin (2001).

turn means that the proportion of cooperators in the group also approaches 1.

If the group is sufficiently large, cooperators conclude that the proportion of cooperators is greater than $1/k$. Once this threshold is reached, cooperators follow their cooperative strategy in the public goods game. The result is that the material payoffs for cooperators increase as the proportion of cooperators grows. Naturally, the remaining defectors also benefit as their opportunities for exploitation increase. The remaining defectors still do individually better than the cooperators, but the rate of defectors in the group decreases as the group grows.[8]

Cooperation turns into a viable strategy. Weak translucency and an appropriate process of group assortation enable cooperative agents to thrive. While the model is extremely simple, it provides a basic mechanism by which cooperative strategies can survive even in hostile one-shot dilemmas. Survival is possible if weak translucency and appropriate social conditions allow agents with "nice" strategies to change the environment they are operating in. Weak translucency, information pooling and group assortation protect cooperators from free-riders.

In the next section I formalise and extend this simple model. In particular, I limit the number of agents, introduce two groups, make some more realistic assumptions regarding the group formation process, and run some simulations.

6.5. A More Realistic Model

Let there be n agents. There are three stages in the game: An information stage, a group formation stage, and a game stage. At the information stage, each agent receives a private signal about the disposition of every other agent. At the group formation stage, each agent is either assigned to group 1 (the "in-group") or group 2 (the "out-group"). The two groups have g_1 and g_2 members respectively, with $g_1 + g_2 = n$. At the game stage, the two groups each play a public goods game, and the agents receive the resulting payoffs.

In order to enhance the clarity of the explanation I begin—against the temporal sequence of the stages—with the game the two groups play.

Game Stage

Once groups 1 and 2 are formed, both groups play a separate public goods game. G_1 is the set of agents in group 1, G_2 the set of agents in group 2. Each agent makes a

[8] One could also think of an exclusion mechanism, forcing members to leave if a majority or a supermajority decides they must. This would root out defectors entirely.

contribution c_i of either 0 or 1 units to the public good fund, such that $c_i \in \{0, 1\}$. The net payoff a_i for each agent is

$$a_i = \begin{cases} k\frac{\sum_{j \in G_1} c_j}{g_1} - c_i & \text{if agent is in group 1} \\ k\frac{\sum_{j \in G_2} c_j}{g_2} - c_i & \text{if agent is in group 2,} \end{cases} \tag{6.1}$$

with $1 < k < 2$. This is a public goods dilemma where contributing is collectively preferable, while defecting is individually preferable.

In monetary terms, agents face an ANPD. There are two types of agents, signified by

$$d_i = \begin{cases} 1 & \text{if cooperator} \\ 0 & \text{if defector} \end{cases} \tag{6.2}$$

for each agent i. Defectors do not have any desire to forgo the pursuit of their own monetary self-interest. Therefore, defectors always defect. Cooperators, by contrast, are committed to cooperate as long as they do not expect to encounter a monetary loss when they cooperate. There are x cooperators and y defectors with $x + y = n$. Cooperators expect a loss when r, the proportion of cooperating agents in the group, is smaller than $1/k$.

Cooperators cooperate as long as they expect that cooperation yields a positive payoff for them. Cooperation yields a positive payoff if and only if $r > 1/k$. If they expect a negative payoff they defect. This implies that cooperators do not maximise material payoff. Usually, cooperators do not know for sure whether $r > 1/k$ holds. This is why they have to infer the size of r from their own information and the behaviour of other agents as explained below.

Information Stage

At the information stage, each agent receives a private signal about the disposition of every other agent, that is whether the other agent is a defector or a cooperator. Let the signal s_{ij} be the signal that agent i receives about agent j. $s_{ij} = 0$ if and only if i receives a signal that j is a defector; $s_{ij} = 1$ if and only if the signal indicates a cooperative disposition. Each agent receives information about all other $n - 1$ agents.

The reliability of this information is given by the conditional probabilities

$$p = Pr(s_{ij} = 0 \mid d_j = 0) = Pr(s_{ij} = 1 \mid d_j = 1) > 0.5, \tag{6.3}$$

for any two agents i, j. The probability to receive a correct signal, p, is thus better than random.

Group Formation Stage

Before the game stage, agents form groups in which they play the game. I have previously assumed that there is an unlimited pool of agents and that the group size approaches infinity. I now replace this idealised assumption with a more realistic setting. In the beginning, both groups are empty. They are formed according to this scheme:

1. One agent (the "founder" of group 1) is chosen at random and is assigned to group 1.
2. One agent i, who is not yet assigned to any group, is randomly chosen.
3. Group 1 decides whether i is admitted by simple majority voting. If admitted, i joins group 1, otherwise group 2. If either of the two groups has reached its maximum size g_1^{max} respectively g_2^{max}, i joins the other group.
4. Move to step 2 until all agents are assigned to either group 1 or group 2.

This scheme assigns all agents either to G_1 (the "in-group"), or G_2 (the "out-group"). It is likely that the rate of cooperators in G_1 is higher than in G_2. This is because the members of G_1 go through a selection process, while the members of G_2 are the remaining or rejected agents.

All agents prefer to be in a group with more cooperators rather than with fewer. Cooperators have this preference because it increases their chances to realise cooperation gains; defectors because their payoff from free-riding increases. In this chapter I assume that all agents vote informatively for or against the admittance of a new member to G_1. This means that agents simply vote according to the private signal they have regarding the candidate.[9]

[9] I discuss problems arising from strategic voting below.

Contribution Decision

At the game stage, the described game is played separately in the two groups. Defectors prefer to defect in all situations. For cooperators, the choice of strategy is not as straightforward. They prefer to cooperate when they expect a rate of more than $1/k$ cooperating agents in their group, implying a positive payoff for the cooperator. However, the cooperator does not know how many cooperators are in the group and how many of them will decide to cooperate. Therefore, cooperators must estimate the rate of cooperating agents.[10]

I consider two different heuristics for performing this estimation. Both assume that agents are boundedly rational. They either use only their own private signals to estimate the number of cooperators in the group, or take the voting behaviour of others into account. This leads to these two estimation procedures:

Private Judgement (PJ). Agents rely only on their own information, i. e. their private signals about the disposition of the other agents, when predicting the behaviour of fellow group members. In that case, a cooperator i cooperates as long as the following inequality holds:

$$1/k < \begin{cases} \dfrac{1 + \sum_{j \in G_1, i \neq j} s_{ij}}{g_1} & \text{if agent is in group 1} \\ \dfrac{1 + \sum_{j \in G_2, i \neq j} s_{ij}}{g_2} & \text{if agent is in group 2.} \end{cases} \tag{6.4}$$

Group Judgement (GJ). The agents take the group formation process into account. If agents believe that the average competence of all agents is greater than 0.5, and if group 1 is large, then members of group 1 expect the proportion of cooperators to be high in their group, while agents in group 2 expect the proportion of cooperators to be low in their group. To simplify, let cooperators cooperate if and only if they are

[10] Teddy Seidenfeld (personal communication) pointed out to me that one could take a different perspective to this question: One could argue that the setting creates act-state dependence. If I plan to cooperate, this is evidence that I am probably in the in-group (with many other cooperators). If I plan to defect, this is evidence that I am in the out-group (with many other defectors). For an evidential decision theorist this implies that it might be rational to cooperate, because of the described act-state dependence. However, for a causal decision theorist this does not follow, because my cooperation decision has no causal influence on the cooperation decision of other players. My sympathy is with the causal decision theorist. I still believe that the dominance principle applies in my model and that a payoff-maximising agent should defect. However, this discussion is beyond the scope of this chapter.

members of group 1.

If agents had full information about the competence of all other agents and the rate of cooperators in the population, they could use Bayesian updating to estimate the rate of cooperators in their group and the implied monetary gain to be expected. However, it is not very realistic to assume that Bayesian updating is a good approximation of human decision making, when it comes to judgements about the cooperative disposition of people. Such judgements are typically seen as judgements about the moral integrity of a person. It is unlikely that individuals have a good idea about their own competence or that of other people regarding integrity judgements. Also, belief in moral integrity has a strongly binary character. We either believe in a person's moral integrity or we don't. We do not think of moral integrity as a probability. In addition, GJ is informationally and computationally much less demanding than full Bayesian updating. Therefore, it is probably a better approximation of human decision making. The following simulations show that GJ is already a very successful strategy. Thus, even though GJ may not be the optimal strategy to play, it does well enough to deliver the desired effect.

Summary of the Model

There are n agents with x cooperators and y defectors ($n = x + y$). In the beginning, cooperators and defectors are not sorted into any groups. They have enough contact with each other to form opinions about each other, i. e. whether they are cooperators or defectors. Agents are translucent, which means they cannot entirely hide their type. Each agent has the probability p (with $p > 0.5$) to correctly recognise the type of another agent. Each agent receives private signals about all other agents. One possible way to think about this signalling is through an initial meeting of all subjects at some socialising activity.

After the information stage, groups form according to the scheme described above. The agents already assigned to group 1 (the "in-group") can use the private signals they have received to decide which agents they vote for in the admittance procedure. I have stipulated that group 1 uses simple majority voting to accept new members. The group formation process is set up such that groups 1 and 2 have maximal sizes g_1^{max} and g_2^{max}.

Both groups separately play a public goods game. The contribution of single agents cannot be observed. For each group, the contributions are multiplied by k and distributed equally among all members of the group. If translucency is weak, it is rather unlikely that cooperators and defectors are perfectly separated. This has an impact

on the cooperation decision in the public goods game. Defectors always defect in this game. Cooperators, however, must decide whether they should cooperate. They are willing to cooperate as long as their payoff is positive, i. e. if a rate of more than $1/k$ fellow group members cooperate. Since cooperators are only boundedly rational, they use a heuristic to decide. They either estimate the rate of cooperation according to their private judgements, or cooperate whenever they are in group 1, hoping that the assortation process has worked well.

The question is how well cooperators and defectors do in such a setting with different parameter values. To answer this question I run computer simulations.

6.6. Simulation

Agents are involved in one-shot public goods games. Since the contributions in the game are not revealed, there is no "shadow of the future" and punishment is impossible.[11] To determine the success of cooperators and defectors, I run simulations of the model. As the model is stochastic, the simulation must be repeated many times to determine the average success of the types "cooperator" and "defector".[12] In all simulations I assume that $k = 1.5$.[13] For simplicity, I also assume that each agent is assigned to exactly one group (implying $g_1 + g_2 = n$), and that there is exactly enough space in groups 1 and 2 to contain all cooperators or defectors respectively (implying $g_1^{max} = g_1 = x$ and $g_2^{max} = g_2 = y$ after the group formation process).

The success of cooperators hinges on three factors:

- the reliability of the private signals p;
- the proportion of cooperators in the population x/n;
- the size of g_1.

The higher p, the more likely it is to distinguish cooperators from defectors, which increases the success of cooperators. Since cooperators are successful when they meet

[11] There is, however, a "shadow of the past" in the form of translucency signals, but this does not change the one-shot character of the game.

[12] Since the interactions are one-shot, averages are not for specific agents, but for types of agents. Each single agent only plays once. However, one could also interpret the rounds as repeated games such that the same agents form new groups in every single round. To maintain the latter interpretation one has to assume that the track record of agents is not known to other agents and that agents play exactly the same strategy in each round. This interpretation does not change the results. The interactions are still one-shot in the sense that there are no sanctions available, but they are not one-shot in the literal sense, because agents play the same game again and again.

[13] I expect the results to be robust for admissable values of k ($1 < k < 2$).

				private judgement		group judgement	
				cooperator	defector	cooperator	defector
n	$x = g_1$	$y = g_2$	p	type	type	type	type
20	10	10	1.0	.41	.07	.41	.07
20	10	10	0.8	.14	.08	.22	.19
20	10	10	0.7	.04	.07	.09	.27
20	10	10	0.6	-.01	.08	-.02	.33
50	25	25	0.8	.30	.07	.36	.10
50	25	25	0.7	.10	.06	.23	.19
50	25	25	0.6	.01	.04	.03	.31
100	50	50	0.8	.40	.05	.43	.05
100	50	50	0.7	.15	.04	.34	.12
100	50	50	0.6	.01	.01	.12	.26
200	100	100	0.6	.01	.01	.24	.19

Table 6.2.: Results I — Average payouts in the public goods game for types cooperators and defectors with variation in n and p.

their own kind, a high proportion of cooperators in the population is also beneficial for cooperators. However, a high rate of cooperators also creates good conditions for defectors, because defectors thrive only if they meet cooperators. Finally, the size of group 1 is positively correlated with the success of group 1 to exclude defectors. If group 1 is large, the voting process to admit new members produces increasingly correct decisions. In addition, if cooperators base their cooperation decision on group judgement, the size of the group also has a strong positive impact on the correctness of cooperation decisions, i. e. cooperators cooperate more often in situations in which it is indeed beneficial to cooperate.

I have tested the model with several parameter values in a computer simulation with 1000 rounds each. Table 6.2 shows the average payoffs for both types, given different parameter values, comparing private judgement (PJ) and group judgement (GJ). The first column states the number of agents n. For simplicity, I assume that the actual and maximal group sizes are $n/2$ (indicated in columns 2 and 3), and that half of the agents are cooperators, half defectors. The fourth column gives the quality of the translucency signals p. Columns 5 and 6 show the average payoff in the public goods game for the types cooperator and defector, given that cooperators use the heuristic "Private Judgement" (PJ) to determine their contribution. Similarly, columns 7 and 8 give the results if cooperators use "Group Judgement" (GJ).

The simulation confirms the predictions regarding the reliability of signals and the

				private judgement		group judgement	
				cooperator	defector	cooperator	defector
n	$x = g_1$	$y = g_2$	p	type	type	type	type
100	10	90	0.7	-.02	.01	-.14	.04
100	20	80	0.7	.02	.01	.05	.08
100	30	70	0.7	.06	.02	.21	.09
100	40	60	0.7	.13	.04	.30	.10
100	60	40	0.7	.16	.06	.37	.14
100	70	30	0.7	.21	.09	.38	.20
100	80	20	0.7	.21	.15	.40	.31
100	90	10	0.7	.22	.30	.41	.61

Table 6.3.: Results II — Average payout in the public goods game for the types cooperators and defectors with variation in the proportion of cooperators.

size of group 1. We can see that higher n and g_1 increase the success of cooperators, given equal levels of p and the same heuristic. We can also see that cooperators do better with greater p. It is striking that cooperators can do better than defectors and receive high positive payoffs, even when the quality of the signal is not very good, as long as the individuals use the heuristic GJ. By contrast, if agents rely only on their private information, they tend to play very conservatively, which avoids exploitation, but also diminishes the returns for cooperators. For instance, if the quality of the translucency signal p is merely 0.6 (i.e. only slightly better than random) and the size of both groups is 100, cooperators (.24) fare better than defectors (.19) with GJ. With PJ, however, cooperators hardly ever cooperate and the payoffs for both types are merely .01 on average.

To test the prediction that the proportion of cooperators in the population has an influence, another set of simulations is run, with results in Table 6.3. The predictions are confirmed for the parameters chosen (other sensible parameters lead to similar results). A higher proportion of cooperators improves the chances for cooperators to interact with their kind. A larger group also increases the epistemic abilities to distinguish cooperators from defectors. For example, if there are 70 cooperators and 30 defectors (and respective group sizes), cooperators (.38) do better than defectors (.20) on average. However, as the proportion of cooperators increases, the remaining defectors benefit as well. When there are only very few defectors left, then defectors do well. Typically, a small number of defectors manages to slip into the in-group, in particular at the beginning of the group formation process, when the advantage of information pooling is weak. If the overall number of defectors is small, even a few successful defectors push

up the average payoff for defectors.

6.7. Discussion

I discuss the results of the model, the simulations, and their implications for the debate on rule-following, morality, and rationality. The focus is on the link between translucency and processes of group formation. Finally, some future challenges and unresolved questions are discussed.

Results of the Simulations

The simulations show that even modest levels of translucency suffice to make cooperators more successful than defectors. As the population size increases, the requirements for the quality of translucency signals decrease. The success of cooperators, however, hinges on two distinct uses of information pooling.

The first use of information pooling occurs when members of group 1 admit new members by simple majority voting. The vote pools the private information available to the agents. If the conditions of the Condorcet Jury Theorem hold, the likelihood of correct decisions approaches 1 for large groups.

The second use of information pooling takes place when cooperators use the group judgement heuristic, which the simulations have shown to be superior. With group judgement, cooperators believe that being in the "assorted" group 1 guarantees a high rate of cooperation in the group. Thus, they use the result of the first information pooling (the vote on member admissions) to decide whether to cooperate. The simulations show that a combination of these two distinct uses of information pooling enables cooperators to be successful.

The success of assortation in situations with weak translucency hinges on the ability of agents to make collective use of their private signals. Cooperators can do well and even better than defectors, but only if they pool their signals. Since it is unlikely that translucency is strong in real-world situations, the ability to form collective judgements about the cooperative disposition of agents is decisive when dilemmas in the form of anonymous public goods games must be resolved.

While it is individually payoff maximising to defect, cooperators do better as a type if they manage to interact with their own kind. Therefore, the commitment to cooperate does not harm the cooperators under suitable conditions. To the contrary: it is to their benefit.

Integrity, Translucency, and Group Formation

For the sake of the model, the existence of cooperative and defective dispositions is simply assumed. In reality, people are not natural cooperators or defectors. To be a reliable cooperator, the agent must be *committed* to cooperation. Cooperators have a high level of *integrity* and stick to their normative commitment to cooperate. Defectors claim that they are cooperative, but lack integrity and true moral commitment. Therefore, to distinguish cooperators from defectors, agents must assess the intentions and motives of others. They judge each other's integrity. This, I argue, is what humans do frequently.

So far the argument has shown that *if* individuals receive signals about the integrity of other individuals *and* there is a favourable social framework to form groups, *then* cooperators can protect themselves from the free-riding behaviour of defectors. The question is: How likely is it that the conjuncts of the antecedent are true? The assumption that individuals have a hunch about the integrity of other individuals is intuitively plausible. Integrity judgements depend not only on how people act, but also on the belief that they act for the sake of virtue. This interest in the right motives fulfils a function: It is a mechanism to ensure cooperation in social dilemmas where other social controls fail. One-shot public goods dilemmas, in which people can free-ride without being observed, are a typical example. The ability to distinguish cooperative individuals of integrity from individuals without integrity provides agents with a major advantage.

Several factors render it likely that the quality of translucency signals is sufficiently good. Firstly, individuals usually interact over longer periods of time and get to know each other well. While the model presented in this chapter does not make any assumptions about how the signals are generated, it is likely that people can form opinions of others in interactions before the public goods game. Secondly, people deliberate on the cooperativeness of other people, exchanging their impressions as to how these others have behaved and whether they have demonstrated integrity. While this might cause problems with the independence assumption in the Condorcet jury theorem on the one hand (see below), it might also increase the quality of signals on the other. Thirdly, people have a vested material interest in recognising other cooperative people and singling out free-riders, creating an incentive to invest resources in the observation of others. I have discussed the empirical evidence in favour of translucency above.

The second conjunct of the antecedent presupposes the existence of a suitable framework which allows cooperators to form a group of their own kind, excluding defectors.

This, I argue, is indeed what we observe in real-life group formation processes. These processes rely on the notion of group identity. Identification, in turn, implies that individuals share a set of norms. Agents shape their environment by choosing to interact with people they like and trust. Political parties, non-governmental organisations, pressure groups, local initiatives, etc. all consist of members who share relevant norms and a feeling of "we-ness". Naturally, members of the same group have more in common with each other, and the likelihood that they cooperate successfully in social dilemma situations is *ceteris paribus* higher (for empirical support see Gächter and Thöni, 2005). In the model described above, the process of group formation is extremely simplified, but it models the basic character of this process: A group is an exclusive social entity. It may choose its members and exclude others. With every in-group comes an out-group. In the model, group 1 is the in-group, actively selecting its members, while group 2 is the out-group constituted by those rejected by group 1.

Future challenges

There are at least three challenges that the current model fails to address: Mimicry, the independence assumption, and the problem of strategic voting.

To be perceived as a cooperator is valuable because it yields higher payoffs—as long as cooperators are able to exclude defectors. This raises the problem of "feigning" or "mimicry". If an agent is able to feign signals of cooperativeness, the agent will be invited to the in-group of cooperators, which this agent then exploits by defection. I do not attempt a systematic treatment of this problem. One option is to argue that certain expressions of integrity cannot be feigned (Frank, 1988, ch. 5). The question is then why evolution has not produced mutants able to feign. A second option is to argue that feigning is computationally difficult. Feigning agents have to lie about their motives in public deliberation. Moreover, feigning agents have to make sure that their shirking behaviour is never observed. With more deliberation and interaction taking place, the difficulty to consistently feign one's behaviour grows exponentially. Therefore, I believe that feigning might be possible in short-term interaction, but requires too much computational effort in long-term interactions. These are two possible approaches to the problem of mimicry, but further work is needed.

The model further rests on the assumption that the translucency signals are independent from each other, given the state of the world. This assumption can be problematic, depending on how the signals are generated. Assume the extreme case that every agent meets every other agent one-by-one and forms a judgement without any communica-

tion with others about these judgements. This guarantees independence. In another extreme case, all agents form their signals about all other agents on the basis of rumours spreading through the group. Then the signals are clearly not independent and information pooling will not have the desired effect. In reality, the independence assumption is probably violated, because agents communicate their views of other agents. Nonetheless, the signals might still be independent enough because agents do not take all rumours at face value, but correct them with their own judgements. The signals may also be at least *independent given the shared evidence* the agents have, such that the quality of the collective decision depends on the probability that the shared evidence indicates the true state of the world (see Dietrich and List, 2004). Here majority judgements are still more realiable than individual judgements, but the probability of a correct majority judgement no longer approaches 1 with increasing group size, but only the probability that the evidence is not misleading.

I have so far assumed that agents reveal their private signals truthfully in the voting process of group formation. This assumption is problematic. Austen-Smith and Banks (1996) show that strategic considerations will often motivate agents to vote against their own signals. The problem arises because voters consider their vote in the light of pivotality. The voter realises that her vote is influential only if it is pivotal. Pivotality implies that the other agents are split between acceptance and rejection, i. e. there is a tie between positive and negative signals without the voter's vote. This in turn leads our voter to update her belief on the disposition of the candidate. For instance, the voter could conclude that almost 50% negative signals are overwhelming evidence for a rejection, ignoring her own private signal (the concrete reasoning depends on several parameters involved, including the voter's prior probabilities over the different states of the world).

Austen-Smith and Banks propose a solution. One can adapt the voting threshold in such a way that the strategically best vote is also the informative vote. But for my model, this solution is impractical: As the group size and the available information changes constantly during the group formation process, one would have to set a different threshold for each vote. While this would yield a process immune to strategic voting, it is entirely unrealistic.

The explanatory attractiveness of a fully strategy-proof process can be questioned with regard to its demanding assumptions. Austen-Smith and Banks require a framework of full information (except, of course, about the private signals) and rationality, which hardly ever obtains in reality. Firstly, it is unlikely that agents have the relevant information to engage in Bayesian updating. They need to know both the competence

level of other agents and the prior probability that an agent is a cooperator. Both parameters are usually not known in real-life settings. Furthermore, it also seems unlikely that agents are motivated to vote strategically in the settings discussed here. If people have arrived at a judgement about the cooperative disposition of a person, they are unlikely to misrepresent this judgement for strategic reasons. People usually stick to their judgements about other people and do not update them easily just because some other people have differing views. Therefore, strategic behaviour in the sense of Austen-Smith and Banks may not even be the limiting case of human behaviour in these processes.

The upshot is that agents are very likely to be boundedly rational and to use simple heuristics when they decide which agents to accept in a group and when to cooperate. They may not use exactly the heuristics analysed in this chapter. But this does not threaten the argument: The point is that cooperators can do well even if their behaviour is not fully rational. The model produces a possibility result that remains valid even when the specific assumptions regarding the heuristics are questioned.

6.8. Conclusion

This chapter tests how viable the commitment to cooperate with other cooperators is in the hostile environment of an anonymous public goods problem. It turns out that cooperation is not only viable but highly successful. This success hinges on three mechanisms: translucency, assortation, and information pooling. Translucency is linked with the concept of integrity. People have a hunch about the integrity of other people. We try to interact with people of integrity because we want to avoid persons who are likely to exploit us when they have the opportunity. Integrity is not only a matter of deeds but also of motives. We need to know whether others are truly committed to cooperation—i. e. whether they are motivated by this commitment rather than instrumental maximisation—before we expose ourselves to possible free-riding.

The results of the models presented in this chapter have some interesting implications for the study of group formation and for the emergence of cooperation. The models show that group member selection according to integrity is a device to secure cooperation gains against the constant threat of free-riding. This conjecture is backed up by the fact that group formation and group identity have a strong influence on human behaviour (see for example Akerlof and Kranton, 2000; Monroe, Hankin and Bukovchik Van Vechten, 2000; Monroe, 1997). The presented models and their interpretation in the light of translucency, information pooling, and integrity open us new routes for a

more analytical analysis of these complex phenomena.

The results in this chapter are important in the light of my analysis of subtle sanctioning mechanisms to gain a better understanding of realistic moral theory. We have seen that norm compliance is often a social dilemma, where individual free-riding is tempting. The puzzle is to explain why compliance with many norms is high, even though it seems to be less attractive than defection. The model in this chapter gives an answer: The compliance dilemma can be solved if cooperators recognise other cooperators. This works particularly well when cooperators pool their information and are able to form groups of cooperative agents.

7. Sort out your Neighbourhood: Prisoner's Dilemmas, Anarchy, and Social Structure

Introduction

In *The Prince*, Niccolò Machiavelli observes that

> "[...] many have pictured republics and principalities which in fact have never been known or seen, because how one lives is so far distant from how one ought to live, that he who neglects what is done for what ought to be done, sooner effects his ruin than his preservation; for a man who wishes to act entirely up to his professions of virtue soon meets with what destroys him among so much that is evil." (Machiavelli, 1958, ch. XV)

Machiavelli sees a gap between "how one ought to live" and "how one lives". The reason for this gap is the desire for one's "preservation" in an environment where enough people are "evil" and will take advantage of those who act virtuously. Therefore, Machiavelli urges the Prince to eschew utopian notions of virtue: "Hence it is necessary for a prince wishing to hold his own to know how to do wrong, and to make use of it or not according to necessity." (Machiavelli, 1958, ch. XV). Virtuous behaviour would be possible if everyone behaved virtuously—but in the real world, "among so much that is evil", this leads to self-destruction. The Prince must learn how to do wrong, and he must "be a fox to discover the snares and a lion to terrify the wolves." (Machiavelli, 1958, ch. XVIII).

In the same vein, Thomas Hobbes describes the State of Nature as a state of war because agents are trapped in the dilemma of pre-emptive measures to secure themselves:

> "Also, because there be some that, taking pleasure in contemplating their own power in the acts of conquest, which they pursue farther than their

> security requires, if others, that otherwise would be glad to be at ease within modest bounds, should not by invasion increase their power, they would not be able, long time, by standing only on their defence, to subsist." (Hobbes, 1996, ch. XIII)

The problem of security is old, yet still of great importance: Failed states, anarchy and civil war all teach us that the provision of security and public order is a social dilemma, and it is very difficult to return to a state of peace once the provision of security has broken down. In recent years, Somalia, Afghanistan, and now Iraq have given us sad examples of what happens when the dilemma cannot be solved. Most people experiencing anarchy or the quasi-anarchy of civil war would wish to return to a peaceful co-existence. However, once the dilemma has them in its grip they are forced to arm themselves and perhaps even fight preemptively for their own protection, making peace ever more difficult.

	cooperate	defect
cooperate	2, 2	- 1, 3
defect	3, - 1	0, 0

Table 7.1.: Game form with payoffs in the structure of a prisoner's dilemma.

The dilemma outlined by Machiavelli and Hobbes is a special case of the problem of cooperation. At its most basic form, the dilemma comes as a 2-person prisoner's dilemma (PD). Table 7.1 shows a game form with payoffs for Row (the row player) and Column (the column player). If we assume that the players are payoff maximisers, then the game form and the players' preferences induce a PD and mutual defection is the strict and only Nash equilibrium. One can interpret Machiavelli's problem as a 2-person PD: It is best for Row if Column acts virtuously (cooperates) and Row exploits him by acting evil (defects). The second best outcome is when both act virtuously (both cooperate), the third best when both choose evil (both defect), the worst when Row acts virtuously (cooperates) and Column exploits him by choosing evil (defects). Column has analogous preferences over outcomes. Defect/defect is the strict Nash equilibrium, and Machiavelli argues that it is quite likely that players end up in this equilibrium. Thus, the prince better prepares himself and practises being evil when it is necessary.

Hobbes's dilemma of security can also be interpreted along these lines. Row and Column can either choose peace (cooperate) or war (defect). Both players order the outcomes as follows: It is best to wage war against an opponent who chooses peace. It is second best when both players choose peace. It is third best when both players

	cooperate	defect
cooperate	3, 3	0, 2
defect	2, 0	1, 1

	cooperate	defect
cooperate	3, 2	0, 3
defect	2, 0	1, 1

Table 7.2.: Assurance Game (left) and a hybrid between assurance game and prisoner's dilemma (right). Numbers are ordinal preferences.

choose war. It is worst to choose peace when the other chooses war. Therefore, rational players will end up in the Nash equilibrium war/war.

Peter Vanderschraaf (2006) argues that Hobbes's dilemma is in fact not a prisoner's dilemma, in contrast to the conventional wisdom. In the quotation above, Hobbes talks of "others, that otherwise would be glad to be at ease within modest bounds". Those "others" do not play a PD, because they have peace/peace as their most preferred outcome. They value peace higher and dislike war more, compared to their more ambitious opponents. It seems that Hobbes (and possibly also Machiavelli) have a situation in mind where some people play a PD, while others play an assurance game. PD players always defect (or choose war, or are evil). Assurance game players are in a more difficult position: They would most prefer to cooperate if the other player cooperates, but they would rather defect if the other defects.

Distinguishing between game form and game clarifies the situation: All individuals face the same game form with payoffs as stated in table 7.1. But they play different games because they have different preferences over the outcomes. The numbers in table 7.2 state preferences, not payoffs. Table 7.2 (left) shows an assurance game. Here both players must prefer the outcome of mutual cooperation. The table on the right is a hybrid between a PD and an assurance game: Row plays an assurance game and would most prefer mutual cooperation. However, Column plays a PD and defects in this game. The tricky question for the "friendlier" assurance game player Row is: Against which type is he playing? Is he in the benign game on the left or in the nasty game on the right?

Vanderschraaf arrives at a pessimistic conclusion, vindicating Machiavelli's and Hobbes's informal argument that peace (or virtue) are not possible when at least some fellow citizens are ambitious or evil. My analysis offers a more optimistic outlook: Under certain conditions the dilemma can be solved if agents can change the social structure of their interaction. My way of modelling the dilemma differs from Vanderschraaf's in several ways. Firstly, I use a dynamic network analysis to model social space. This allows agents to influence with whom they interact by changing the social relations they have. Secondly, I see the dilemma as a n-person public goods problem, while Vanderschraaf

uses 2-person games.

Networks offer a more realistic account of how social interactions evolve over time. Agents choose to end relations with people who tend to defect, while trying to build new social ties to cooperative agents. Dynamic networks capture these evolving social structures. The advantage of using n-person games to analyse Machiavelli's and Hobbes's dilemma is that they offer a more realistic model of social interactions. 2-person games, in particular prisoner's dilemmas, are omnipresent in the literature, but they are rare in reality. The agent in the State of Nature does not have one stand-off with one person at a time. The problem is rather that there are many persons and all of them can hit the agent from behind at any point in time. n-person games are also more relevant in less archaic settings: Public goods provision and collective action problems are n-person dilemmas. In fact, almost all social dilemmas involve more than 2 persons, and the 2-person case is captured as a special case within the more general n-person setting.

The chapter starts with a short discussion of why I choose to work with computer simulations rather than deductive game-theoretical analyses. Section 2 gives an introductory example of agents playing 2-person prisoner's dilemmas on a dynamic network. Section 3 introduces the core model: n-person prisoner's dilemmas on dynamic networks. This is followed by an exploration of how robust the derived results are. The chapter concludes with some thoughts on how models with dynamic networks can be used to analyse a wide range of social phenomena.

7.1. Dynamic Networks and Computer Simulations

Until recently researchers have paid little attention to the influence of spatial structure on the emergence of cooperation. Notable exceptions are Nowak, Bonhoeffer and May (1994) and Alexander (2003*b*, 2007), pointing out that spatial arrangements can have an important impact. Nevertheless, most research has focused either on homogeneous populations with no spatial restrictions for the interactions of agents, or on lattices with interactions restrained to neighbours on a grid. In the first wave of literature (Lieberman, Hauert and Nowak, 2005; Ohtsuki et al., 2006; Ohtsuki and Nowak, 2006*a*; Santos and Pacheco, 2006; Ohtsuki and Nowak, 2006*b*, and further references in these papers) the structure of the network is static. The network structure influences the agents' behaviour and payoffs, but agents are not able to change the structure of the network. This chapter, by contrast, implements dynamic network structures (cf. Pacheco, Traulsen and Nowak, 2006*a*). Agents can influence the agents they have contact with

and thereby shape their neighbourhood. This mirrors the nature of social structures in reality: We have some, but not complete control over the set of people we interact with. We can cut ties with those who cheat us and establish ties with those who seem trustworthy. Such networks can be of a professional (trade networks, academic collaborations, etc.) or a private nature (networks of acquaintance, social networks in virtual worlds, etc.).

Social space has an important function for the emergence of cooperation. It regulates who interacts with whom, and it provides opportunities for agents to change their interaction partners. One typical way to implement the notion of space is to use grids or—less technically—checkerboards. Each agent inhabits one field on a checkerboard and has a limited number of neighbours. There are different ways to define a neighbourhood, for example: all fields to the top, bottom, left, and right of the agent, or all fields adjacent to the agent, including diagonally adjacent fields, etc.. It is sometimes assumed that agents can move along the checkerboard, thereby changing their spatial position. The disadvantage of modeling social space as a checkerboard is its rigidity: Every field on the board has a fixed number of neighbours in its immediate local neighbourhood. Real social networks look different: Firstly, agents can differ in their number of social contacts; secondly, these contacts are not necessarily local (think of online communities); and thirdly, real agents have the chance to influence the network structure by making and breaking social relations. To incorporate these properties of real social networks, I model social space as a graph.

A graph consists of vertices and edges. When drawing a graph, vertices are represented as points, and edges as lines connecting these points. Each edge connects two vertices. I take a vertex to represent an agent, and an edge to represent a social relation between two agents. Connected agents are neighbours, which implies that they interact with each other. The network in the model is dynamic. It changes its structure because agents can choose to delete edges and new edges are created. This represents the fact that agents have a choice with whom they have social relations.

Analytical solutions to repeated games on dynamic networks are difficult to find. The space of possible strategies is enormous and the interaction between network structure and game strategies is difficult to capture.[1] If there is a large number of agents and a large number of rounds, it is almost impossible to derive an extended game form and "solve" the game. In any case, it is quite implausible to assume that agents have

[1] Even in Axelrod's (1984) computer simulations of iterated prisoner's dilemmas the number of logically possible strategies is infinite (if the number of rounds has no fixed upper limit). The number of sensible strategies (the standard of what is sensible is of course open for debate) is much smaller, but still large as the number of submitted strategies in these competitions shows.

full information on all outcomes of games played on the network. Also, the complex dynamics of repeated plays in networks, especially if n-person games are played, are difficult to capture analytically. The upshot is that an analytical solution to these complex games is practically impossible and would have to rest on knowledge and rationality assumptions which would render the model unrealistic. Nevertheless, these games can be analysed in greater detail. For complex and dynamic games we need to replace deductive analysis with computational modeling.

The options to model the emergence of cooperation on dynamic network structures are endless. This requires us to tread carefully. Since the design of the model is underdetermined by the research question, there is a danger to produce artifacts. The richness of real-life social interactions supplies us with many ideas for intuitively plausible models, and with many possible heuristics agents could use. In these cases simplicity is a good guiding principle (see chapter 4). The point should not be to represent all subtleties of real life in the model. Rather the model should capture the fundamental mechanisms of cooperation on dynamic networks.

7.2. 2-Person PDs on Dynamic Networks

I start with a very simple model. In the beginning, agents are situated in a social network, with random social structure (see figure 7.1, panel a). Each pair of connected agents plays a 2-person game form with payoffs as stated in table 7.1. For payoff maximising agents these payoffs constitute a prisoner's dilemma.[2] However, not all agents are immediate payoff maximisers in this game: Cooperators always cooperate, even though cooperation is not a Nash equilibrium for payoff maximisers. Defectors, by contrast, always defect.[3] Defectors always do better than cooperators in each single game in material terms. However, agents are allowed to create and sever ties. A cooperator can try to sever ties with defectors and connect to a new agent, hoping that the new agent is a cooperator. Defectors will do the same: Both types of agents want to play with cooperators and want to avoid defectors. In this model, dynamic reinforcement of successful interactions leads to the success of cooperators.

More technically speaking, a network is represented by a graph with n vertices and

[2]For simplicity, I will occasionally call a game with the *monetary payoffs* of a prisoner's dilemma a prisoner's dilemma in this chapter, even though cooperators do not play a prisoner's dilemma in their own perception, all things considered.

[3]Cooperators do not play a PD but a cooperation game because a PD is *defined* such that agents always prefer to defect. However, both cooperators and defectors face a payoff structure typical for a prisoner's dilemma.

k edges. Let the edges be non-directional. Self-loops (a vertex connected by an edge to itself) are ruled out. Each vertex represents an agent. Vertices can be of two types: Cooperators (C) and defectors (D). The edges represent interaction relations between agents such that two connected agents interact with each other in each round of the game. In the beginning, the edges randomly connect vertices.[4] The type of each vertex is also determined at random with the condition that there be x cooperators and y defectors. It is possible to have multiple edges between two vertices.

In each round, every pair of agents connected by an edge plays a PD (in terms of monetary payoff) with each other.[5] Cooperators always cooperate, defectors always defect. The payoffs for the PD are stated in table 7.1. After playing the game agents can choose to delete one of "their" edges, i.e. they can choose to delete one of the ties[6] connecting them to other agents. They can also choose not to delete any edge. This means if an agent i has l edges, i has $l + 1$ alternatives: Delete one of the l edges, or delete no edge.

In the first simulation I assume that cooperators are *zealous*, this means they sever ties with defectors whenever they can. Defectors are *inert*, i. e. they do not change their connections to other agents, because they benefit from having ties with cooperators and are not harmed by other defectors. After all agents had the option to delete one edge, the number of deleted edges is replaced by new random edges.[7] This procedure is repeated for many rounds.

Figure 7.1 shows the effect of repeated play. White vertices represent cooperators, black defectors. In the beginning (panel a) players are randomly matched. After 100 rounds (panel b), the network has changed its structure. Cooperators are only connected to other cooperators, defectors only to defectors. The situation depicted in 7.1b is stable with the strategies described. Neither cooperators nor defectors have reason to sever any ties, given the strategies *zealous* for cooperators and *inert* for defectors. Since new ties are only established when old ties are deleted, no change in the network structure takes place once cooperators and defectors are completely separated. The payoffs for defectors are higher than for cooperators in the beginning, but separation of the two soon puts cooperators in a better position. In figure 7.1b defectors receive zero payoff, while cooperators receive payoff 2 for each tie to another

[4]Note that the graph is not a complete graph, i. e. typically many pairs of vertices are not directly connected.

[5]If there are multiple edges between two agents, they play the game as often as there are edges between them. Multiple edges can be interpreted as representing a particularly intensive interaction.

[6]I use the terms "edge" and "tie" interchangeably throughout the chapter.

[7]For simplicity, I assume that the new random edge can also be the old, deleted edge. This, of course, is very unlikely for a sufficiently large network.

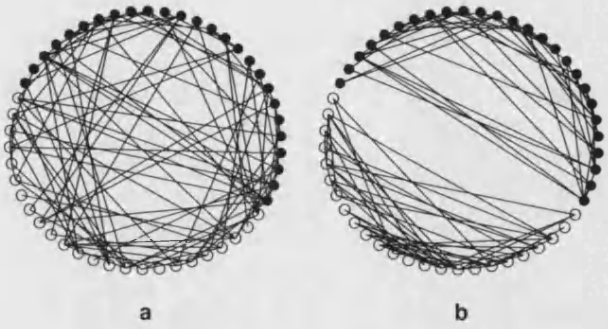

Figure 7.1.: Complete assortation for a 2-person PD with cooperators severing ties to defectors. Cooperators are white, defectors black. (a) is the initial setting with 50 vertices and 100 edges, (b) the network structure after 100 rounds.

cooperator.

With a slight modification the effect becomes even more dramatic. Assume that defectors are *zealous*, too, i. e. they sever ties to other defectors if they can (to motivate this, assume that mutual defection leads to a negative payoff in this game). Figure 7.2 shows the result. Since defectors no longer keep their edges to other defectors, the cooperators camp gets all edges in the network, and defectors are isolated with no ties to other agents.

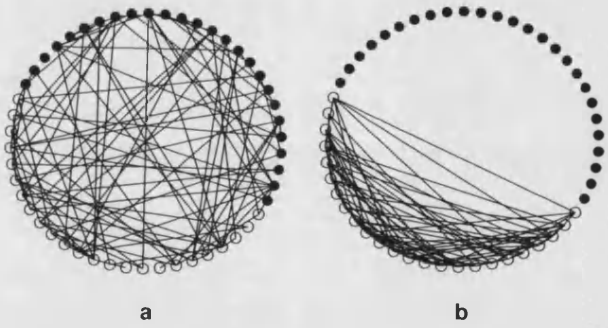

Figure 7.2.: Complete assortation for a 2-person PD with cooperators *and* defectors severing ties to defectors. (a) is the initial setting with 50 vertices and 100 edges, (b) the network structure after 100 rounds.

Even though the model is simple, it already provides some useful insights. Firstly, it shows that network dynamics are a powerful mechanism to enforce cooperation. Without network dynamics, the best cooperators can do is to play a conditional strategy

such as Axelrod's TIT-FOR-TAT. Such strategies 'punish' defectors with defection. These punishments are costly. By contrast, moving away is a cheap but highly effective punishment, because it imposes future losses on the defector, while giving the punisher a good chance to increase payoffs by finding a better partner. Secondly, despite its simplicity, the model gives us a good idea of how some social interactions work. Buyer-seller relations often resemble 2-person PDs: The buyer can refuse to pay, the seller can refuse to send the goods (or send faulty goods). If an agent finds that her business partner has cheated her, she stops dealing with him and finds new partners. In this way business networks of reliable traders emerge even though other enforcement mechanisms are missing (proceeding against someone in a different country is often not worth the effort). However, this will only work if both sides expect future interactions. Without a shadow of the future, neither side has an incentive to cooperate. The game can also be seen as a highly idealised version of Machiavelli's virtue game: The virtuous cooperators lose against the evil defectors. However, in this game the virtuous can move away and exclude the evil-doers.

7.3. n-Person Prisoner's Dilemmas on Networks

More interesting constellations arise when cutting ties to defectors is not that easy. Realistic models comprise n-person PDs with anonymous contributions. This means that agents only know how many players play and how the outcome differs from the ideal outcome of universal cooperation. Rather than cutting specific ties to defectors, cooperators can only try to gradually "move away" if they are caught in a neighbourhood with high levels of defection. Surprisingly, cooperators can do well even in public goods games with anonymous contributions.

Many real social dilemmas involve more than two persons. The paradigmatic cases are collective action and public goods problems. I see Hobbes's security dilemma and Machiavelli's problem of the virtuous as a public goods game: Peace or mutual virtue are public goods. However, for each single individual it is better to free-ride and go to war or be evil when this promises higher individual payoffs. Thus, the agents collectively fail to provide the public good of peace or virtue because they are seeking their own advantage. I argue that social space and the exclusion of the mischievous individuals are crucial elements to a solution of this dilemma.

When many persons are involved, it is often not known who has cooperated or defected. People can often get away with free-riding, because there are no effective ways to monitor behaviour. The more anonymous interactions are, the easier free-riding gets.

For instance, it is often convenient to dump one's rubbish into the street in a moonless night (defect), rather than separating it and carrying it to the next recycling center (cooperate). If no one is watching, or if people do not know each other well enough to identify offenders (think of large anonymous blocks of flats), free-riding remains undetected or unpunished.

However, people do observe how well the production of public goods is going on the aggregate level. Even though it is difficult to find out *who* is littering on the street, it is easy enough to see the dirt. This is the idea for the next model. Again, there are two types of agents, cooperators (or contributors) and defectors (free-riders). Agents are again represented by vertices in a network. In each round, each agent i plays an n-person prisoner's dilemma with all of i's directly connected neighbours.[8] The set of neighbours of i *including* i is denoted H_i, and $|H_i|$ is the number of agents in i's neighbourhood including i. Each agent i makes a contribution $c_i \in \{0, 1\}$. The net payoff p_m^i for each participant m in the game initiated by i is defined as

$$p_m^i = \begin{cases} r\frac{\sum_{k \in H_i} c_k}{|H_i|} - c_m & \text{for } |H_i| \geq 2 \\ \\ 0 & \text{for } |H_i| < 2. \end{cases} \tag{7.1}$$

r is a parameter with $1 < r < 2$. For convenience, I assume $r = 1.5$ unless otherwise stated. If all agents contribute, each agent receives a net payoff $r - 1$, provided there are at least 2 players. Defection is the strictly dominant strategy for payoff-maximisers. However, cooperation can be a viable strategy if cooperators manage to play the game only or primarily with other cooperators.

Figure 7.3 gives an example. Agent m has edges with agents a, b, and c, who again have edges with other agents. Remember that all agents play the public goods game with their neighbours. Here m plays one game with $\{a, b, c\}$, but m also participates in the games initiated by all direct neighbours. Therefore, m participates in *four* public goods games.

In general terms, the payoff p_m for an agent m is determined by adding the payoffs from all the public good games m is playing. This results in

$$p_m = \sum_{i \in H_m} p_m^i. \tag{7.2}$$

[8] It is possible that an agent is connected to the same agent by multiple edges. In this case, this neighbour counts as multiple neighbours. One could say that the multiply connected agent has a higher stake in the game.

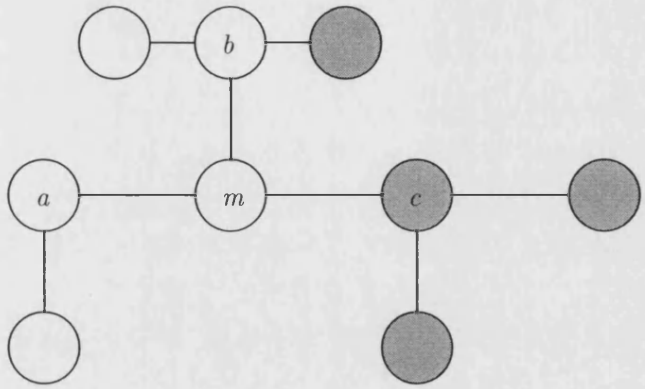

Figure 7.3.: A network constellation. Grey circles are defectors, white circles cooperators.

After playing in the constellation as shown in figure 7.3, the following information is available to agent m:

1. Agent m knows the number of cooperators c_m and the number of defectors d_m in his neighbourhood H_{-m} (H_{-m} is the neighbourhood of m without agent m himself, that is $H_m/\{m\}$).

2. Agent m knows the respective numbers of defectors d_a, d_b, and d_c and cooperators c_a, c_b, and c_c in the neighbourhoods H_a, H_b, and H_c.

In the simulation agents can influence their network by severing ties. Severed ties are replaced by new random ties in the network. Thus, agents are able to gradually change their neighbourhood when the level of cooperation is unsatisfactory. The question is whether cooperators manage to find cooperative neighbourhoods given that they cannot identify defectors directly.[9] The process of the game is summarised in figure 7.4.

Agents need a criterion to decide when to sever ties with other agents. Rational agents should try to determine this criterion by calculating the expected utility gain or loss from severing a tie. To undertake this task it is necessary to understand the information available to an agent after the playing stage is over and before the network change stage begins. The agent does not receive information about who among the neighbours and the neighbours's neighbours is a defector or cooperator, unless this can be inferred from the information described. In the example, m knows from the game

[9]Except for the special case that a cooperator has only one neighbour, i.e. $n = 2$. In this case cooperation or defection can be detected.

1. Create non-directed random graph S with exactly n vertices (representing agents) and k edges, without self-loops.
2. For each agent i (represented by a vertex) play a public goods game as defined in formula (7.1) with all agents directly connected to i and change budgets of all agents accordingly.
3. In random order, each agent can choose to sever one of the edges connecting the agent to another agent (if the agent is unconnected, nothing happens).
4. The number of deleted edges is replaced with new edges between randomly selected agents (no self-loops allowed).
5. Go to step 2, or stop if the maximum number of rounds is reached.

Figure 7.4.: The game procedure for the n-person public goods game on a graph.

initiated by m that there is one defector in the immediate neighbourhood. From the other games m knows that a and a's neighbour are cooperators. He infers that in the set of b and b's neighbours without m (denotes as $H_{b,-m}$) are 1 defector and 2 cooperators. In addition, m knows that c and his two neighbours are defectors. This in turn also leads to the conclusion that b is a cooperator.

The example also demonstrates that agent m is not only interested in the type of the immediate neighbours. Since m is involved in 4 games (one initiated by himself, three others initiated by a, b, and c), m cares about the types of his second degree neighbours as well. In the example, m should delete the edge with agent c, as cooperating with c and her neighbours leads to negative payoffs. However, in general it is not trivial to see whether an agent should sever ties and to which neighbour. To answer this question it is necessary to calculate the expected[10] payoff change from round t to round $t+1$ caused by severing an edge $\{m, x\}$. When severing a tie to an agent $x \in H_{-m}$ the expected payoff changes in two ways. Firstly, x no longer participates in the game initiated by m. Secondly, m no longer participates in the game initiated by x. The expected payoff change $\Delta E(p_m)$ is

$$\Delta E(p_m) = [p^m_{m,-x} - p^m_m] - p^x_m \tag{7.3}$$

with $p^m_{m,-x}$ being the payoff m receives from the game initiated by m but played without agent x. The value of the term $p^m_{m,-x} - p^m_m$ depends on whether x is a cooperator

[10]The change is expected because it rests on the assumption that everything else remains equal, which is not necessarily the case. m's neighbours can also delete edges.

or a defector.[11] In most cases m does not know the types of his neighbours. However m is able to estimate the probabilities that a neighbour is a defector based on the outcomes of previous games. The complexity of these calculations depends on the sophistication of the agents. With perfect memory and assuming that agents are able to start with suitable prior beliefs, a Bayesian treatment would be possible. A Bayesian agent uses all information that becomes available during the course of the game and updates her beliefs about the types of all other agents. In this chapter I restrict myself to much simpler and arguably more realistic heuristics.

There are three reasons why a Bayesian calculation of the expected utility change is of little practical relevance. Firstly, even under a strong "as if" assumption for the behaviour of agents, it is unlikely that agents individually or on average behave like perfectly rational Bayesian agents, as the level of computational effort is enormous.[12] Agents have incomplete information and limited cognitive abilities. Therefore they have to use simplifying heuristics to make decisions. Secondly, it is more interesting to show that even rather unsophisticated agents can reach structures where cooperators keep defectors in check. In realistic settings, agents use simple heuristics, and it is of little interpretative interest to model agents as much more sophisticated than they actually are. Thirdly, the choice that maximises the expected utility change *for the next round* is not necessarily the best choice in the context of the repeated game. In Axelrod's computer tournaments, TIT-FOR-TAT turned out to be a rather successful strategy, even though defection is the best response in every single round. However, analytical solutions for such complex repeated games are currently not available.

To explore the ensuing dynamics, it is a good idea to run computer simulations with some plausible strategies. The estimation heuristic underlying all my strategies assumes that agents base their network choice exclusively on the outcome of the games they have played in the last round. The rule is: if an agent wants to sever a tie, the agent severs the tie to the agent whose neighbourhood had the highest rate of defection in the game initiated by that agent in the previous round. In the case of figure 7.3, this choice is obvious: In the game initiated by m himself, m infers that there is one defector in the immediate neighbourhood. When playing the game initiated by a, b, and c, m realises that c's neighbourhood has the highest rate of defection (here, without m it is

[11] Note that these considerations assume that the rest of the network remains stable. If an agent took the dynamic character of the network into account, he would still aim to sever ties to defectors, but the incentive to do so might be even stronger, given that defectors are more likely to have defectors as neighbours.

[12] Note that if the game is played over many rounds with a limited number of agents, the evidence gathered in the current game should lead to a revision of all earlier reasoning processes based on earlier evidence. This is computationally very demanding.

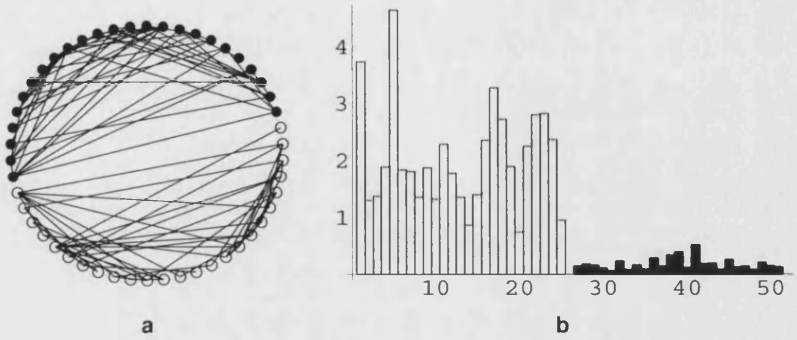

Figure 7.5.: Panel a shows complete assortation for a n-person PD with defectors playing *inert* and cooperators playing *zealous* after 100 rounds. There are 25 cooperators and 25 defectors, connected by 100 edges. Panel b shows the average payoffs per round. White bars are for cooperators, black bars for defectors.

100%). m infers that c must be the defector. But even if no certain inference can be drawn, it makes sense to assume that the neighbour with the highest defection rate in his neighbourhood is the most likely defector in m's neighbourhood.[13]

To understand the basic dynamics of the game, I begin with two simple strategies already familiar from the 2-person prisoner's dilemma. The *zealous* strategy means that agents always sever the tie to the agent with the highest rate of defection, as long as there is a neighbour with a non-zero rate of defection in the neighbourhood. The *inert* strategy means that agents never sever a tie to another agent. From a myopic perspective (looking only one round ahead), defectors should play *inert*, since they are never harmed by any tie to other agents and severing ties reduces their chances to exploit cooperators. Cooperators, by contrast, should play a less tolerant strategy, and *zealous* is the most extreme version of a strategy trying to get rid of ties to defectors.

Figure 7.5 shows a typical result when 25 *zealous* cooperators play against 25 *inert* defectors, connected by 100 edges, over 100 rounds. Figure 7.5a reveals that cooperators and defectors are completely separated after 100 rounds, and given the strategies *zealous* and *inert* this is a stable state, i. e. the network will not change any further. We can see in figure 7.5b that cooperators are doing very well, while defectors have low average payoffs. Since defectors do not have any connections, and the network is in a stable state, their payoffs in all further rounds will be 0. The assortation procedure has led to

[13] Bear in mind, though, that this reasoning only makes sense if m can remember nothing but the last round.

mutually beneficial ties between cooperators, while the ties between defectors do not benefit the defectors. I ran this simulation 50 times with different random networks as initial setting, and in each simulation complete separation was reached after 100 rounds. The average payoff for cooperators was 2.12, for defectors 0.15.

Defectors "lose" this game because they always end up without any connection to cooperators, that is with zero payoff for all rounds after the assortation is complete and the network is in stable state. The myopic *inert* strategy led to the complete separation of cooperators from defectors. However, if defectors adopt a less tolerant strategy they might be able to avoid a network stable state with complete separation of cooperators and defectors. The question is: If defectors are able to successfully coordinate the strategy they play, which strategy should they choose? I run some simulations to answer this question.

Can defectors have a better strategy than *inert* against *zealous*? To describe the development of the network, it is useful to distinguish three different types of edges. CC edges connect cooperator to cooperator, DD defector to defector, and CD cooperator to defector (and vice versa). Thinking about the dynamics of the game analysed so far, it is obvious that defectors should sever DD edges because this creates a chance for new CD edges, i. e. opportunities for the exploitation of cooperators. Therefore, defectors should not accept all connections to fellow defectors. Rather they should keep the dynamics in the network going and try to avoid settlement into a stable state with full assortation. To do this, defectors must delete some or all connections with other defectors. I run simulations where defectors do not accept defectors in their neighbourhood to test this intuition.

I begin with the assumption that all cooperators and defectors play *zealous*. However, with the given parameters (25 cooperators, 25 defectors, 100 edges, 100 rounds) a stable state with complete separation still emerges in all run simulations.

Figure 7.6 shows a typical result after 100 rounds. The *zealous* strategy has not only led to a complete separation of defectors and cooperators, it has also left all defectors without any connection to other agents. The average payoffs for cooperators are still much better than for defectors, but the results for defectors have improved with *zealous*, indicating that *zealous* at least delays the settlement into complete separation. I ran this simulation 50 times (with different random networks as starting points), and in all 50 simulations complete separation obtained after 100 rounds. The average payoff for cooperators per round was 3.61, for defectors 0.58.

These results are a powerful demonstration that cooperation can fare very well when cooperators have a chance to cluster. Note that the situation is very hostile to coopera-

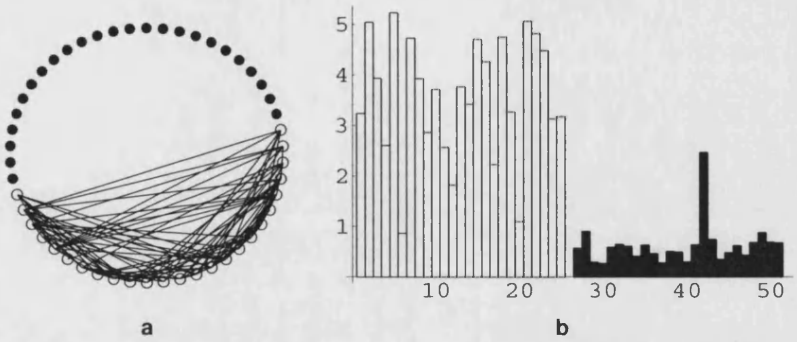

Figure 7.6.: Complete assortation if both cooperators and defectors play *zealous* with 25 cooperators, 25 defectors and 100 edges. Panel b shows that after 100 rounds defectors have far smaller payoffs compared to cooperators.

tion. Cooperative strategies are usually dominated by defection in repeated anonymous n-person prisoner's dilemmas.[14] By providing agents with very moderate levels of information, and—crucially—with the option to shape the environment of the game, it is possible for cooperators to cluster and do well.

The results obtained so far demonstrate that a process of assortation is feasible under specific sets of parameters. Do the specific results hold more generally? Until now the number of agents was assumed to be small, and the number of edges was limited. Also, it would be important to see how the model behaves if the ratio of cooperators to defectors is changed. I turn to these questions in the next section.

7.4. Parameter Variations and Robustness

To assess the relevance of my model, it is important to show the robustness of its behaviour with different parameter values. The problem is that a full exploration of the parameter space is unfeasible, given the restrictions in computing power and the difficulty to derive analytical results for dynamic models. Nonetheless, it is possible to consider at least some sensible parameter constellations to gain a better understanding of the behaviour of the model. I begin with variations in the ratio of defectors. I also explore settings with larger networks and networks with increased interconnectivity.

When there are more cooperators than defectors, the network dynamic still leads to complete separation. With 40 cooperators and 10 defectors, both playing *zealous*,

[14] The Folk theorem shows that any strategy can theoretically constitute an equilibrium in infinite n-person prisoner's dilemmas if the choices of players are public, but in the anonymous situation the Folk theorem does not apply.

and 100 edges in the network, a stable state of complete separation occured in all 50 simulations after less than 100 rounds. Cooperators fared well with an average payoff per round of 2.95, cooperators badly (0.13). What happens when there is a rather small group of cooperators playing against a large group of defectors? In a simulation with 10 cooperators and 40 defectors, with both types of agents playing *zealous*, no separation occurs after 100 rounds. I increased the number of rounds to 1000, but the network still does not settle into a stable state of separation. In 20 simulations, no stable states emerged and cooperators experienced negative payoffs (-1.98), while defectors did well (1.60). Defectors are successful in keeping the dynamics going. Figure 7.7 shows some constellations after 1000 rounds. One can clearly see that cooperators manage to cluster. However, there are still many defectors connected to cooperators, and exploitation of cooperators is widespread. It is also interesting to see that defectors tend to connect with many cooperators. One can interpret this as a "camouflage" effect. A defector connected to many cooperators is less likely to be identified as a defector.

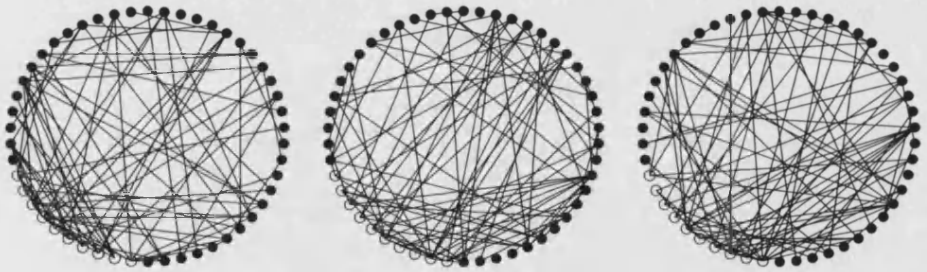

Figure 7.7.: Network constellations after 1000 rounds with 10 cooperators and 40 defectors and 100 edges. Both cooperators and defectors play *zealous*.

Larger networks do not substantively differ in their behaviour from the results observed so far. 100 cooperators and 100 defectors, linked by 400 edges, with both types playing *zealous*, behave almost identical to the smaller model: In 20 simulations over 100 rounds, a complete separation was always reached and the average payoff for cooperators was 3.72, compared to 0.52 for defectors.

A higher number of edges can pose a problem for cooperators. In a graph with high connectivity, it is more difficult to distinguish between cooperators and defectors. This is bad for cooperators and good for defectors. Simulations suggests that it takes more time for the model to settle into a stable state of separation. With 25 cooperators, 25 defectors, 200 edges, and 100 rounds, with both types playing *zealous*, complete separation is usually not reached and defectors (average payoff for 50 simulations: 5.00)

do much better than cooperators (0.77). However, when increasing the number of rounds to 1000, I found separation in 19 simulations out of 20. Cooperators had an average payoff of 5.52, defectors 1.94. Longer play turns around the results to the advantage of cooperators, but the stronger interconnectedness of the network slows down the process of separation.

coop (x)	def (y)	strategy coop, def	edges (k)	rounds	simulations	% stable state	payoff coop	payoff def
25	25	zealous, inert	100	100	50	100	2.12	0.15
25	25	zealous, zealous	100	100	50	100	3.61	0.58
40	10	zealous, zealous	100	100	50	100	2.95	0.13
10	40	zealous, zealous	100	1000	20	0	−1.98	1.60
100	100	zealous, zealous	400	100	20	100	3.72	0.52
25	25	zealous, zealous	200	100	50	8	0.77	5.00
25	25	zealous, zealous	200	1000	20	95	5.52	1.94

Table 7.3.: A summary of all simulation results

Table 7.3 gives a summary of all my simulations. Taking stock, the model displays robustness against variations in line with the theoretical predictions. A complete and stable separation of cooperators from defectors is independent of the number of agents. However, if the rate of cooperators is low, separation does not occur. Also, a higher connectedness of the network slows down the process towards a stable state of separation.

7.5. Some Analytical Considerations

The dynamic of the network model does not lend itself easily to an analytical treatment. Nonetheless, I want to make some analytical observations, without capturing the model completely. The variables e_{CC}, e_{DD}, and e_{CD} denote the number of each edge type in the network, with $e = e_{CC} + e_{DD} + e_{CD}$. The tuple $(e_{CC}, e_{DD}, e_{CD})^t$ states the number of each edge type at time t. Each round is equivalent to one time step.

In each round, as long as the network has not reached a stable state, a number of edges is deleted. This number of edges is then replaced by new edges as explained above. While the overall number of edges is preserved, the composition of edge types changes from $(e_{CC}, e_{DD}, e_{CD})^t$ to $(e_{CC}, e_{DD}, e_{CD})^{t+1}$. Most compositions are *transient* states in the sense that the network is not in a stable state and will change again in the next round. But there are also *absorbing* states where the network dynamic comes to a

halt. Which states are absorbing depends on the strategies played. For instance, when both cooperators and defectors play *zealous*, the only absorbing state is $(e, 0, 0)$. i. e. the state where only CC edges exist. This is because in all other states either cooperators or defectors would still delete edges and keep the dynamic going. When cooperators play *zealous* and defectors play *inert*, all states with zero CD edges are absorbing.

For a thorough analytical analysis of the dynamic it would be necessary to determine a transition matrix that would give the probability for all possible transitions between states. This would then allow us to calculate the probabilities of ending up in different absorbing states. But this approach runs into a problem: The probabilities for the deletion of edges depends not only on specific assumptions on how agents process the information available to them, but also on the specific topology of the network, because the network topology determines which information is available to the agents. Therefore, a thorough analytical approach seems out of reach and simulations are the appropriate tool to analyse the model.

While the deletion of edges poses difficulties, it is relatively straightforward to determine the probability distribution for the type composition of new edges. Let there be x cooperators and y defectors. Remember that new edges are created between randomly chosen pairs of vertices in the graph and that multiple edges between the same agents are possible. The overall number of distinct possible edges is

$$E_{Total} = \binom{x+y}{2} = \frac{(x+y)(x+y-1)}{2}. \tag{7.4}$$

In a similar vein I calculate the number of possible edges between two cooperators (E_{CC}), between two defectors (E_{DD}), and between one cooperator and one defector (E_{CD}):

$$E_{CC} = \binom{x}{2} = \frac{x(x-1)}{2}, \tag{7.5}$$

$$E_{DD} = \binom{y}{2} = \frac{y(y-1)}{2}, \tag{7.6}$$

$$E_{CD} = E_{Total} - E_{CC} - E_{DD} = \frac{(x+y)(x+y-1)}{2} - \frac{x(x-1)}{2} - \frac{y(y-1)}{2} = xy. \tag{7.7}$$

Then the probability that a CC edge is created is

$$Pr(*\text{CC}) = \frac{E_{CC}}{E_{Total}} = \frac{x(x-1)}{(x+y)(x+y-1)}. \tag{7.8}$$

Analogously, the probability that a CD edge is created is

$$Pr(*\text{CD}) = \frac{E_{CD}}{E_{Total}} = \frac{2xy}{(x+y)(x+y-1)}, \tag{7.9}$$

and for a DD edge it is

$$Pr(*\text{DD}) = \frac{E_{DD}}{E_{Total}} = \frac{y(y-1)}{(x+y)(x+y-1)}. \tag{7.10}$$

Assume that s edges have to be replaced. Let $(e^*_{CC}, e^*_{CD}, e^*_{DD})$ be the number of new edges of each type (with $s = e^*_{CC} + e^*_{CD} + e^*_{DD}$). The probability for the compositions of s is multinomially distributed:

$$Pr(e^*_{CC} = z_1, e^*_{CD} = z_2, e^*_{DD} = z_3) = \frac{s!}{z_1! \cdot z_2! \cdot z_3!} \cdot Pr\,(*CC)^{z_1} \cdot Pr\,(*CD)^{z_2} \cdot Pr\,(*DD)^{z_3}, \tag{7.11}$$

where $z_1 + z_2 + z_3 = s$ and $Pr(*CC) + Pr(*CD) + Pr(*DD) = 1$.

As pointed out above, this partial result does not allow us to describe the transition dynamic of the network completely. But some observations can still be made. Firstly, with the strategies for cooperators as described above (both *zealous*, or cooperators *zealous* and defectors *inert*) the network will always reach a stable state in infinite time. I do not offer a a formal proof for this conjecture, but the gist of the argument can easily be seen: There is a small non-zero probability for a transition path from any transient state to the absorbing state(s). With infinite play, the network will eventually randomly end up in the stable absorbing state. This process is similar to neutral drift in evolutionary dynamics (Nowak, 2006*a*, p. 97). Secondly, the fact that many simulations end up in stable states very fast cannot be explained by neutral drift. The reason for these fast settlements is that cooperators delete CD edges more often and CC edges less often than random, thereby pushing the network toward an absorbing state. The success of cooperators to reach a complete separation from defectors depends on their ability to work towards transitions that are beneficial to them.

7.6. Conclusion

The agents in this model operate in a setting that is usually rather hostile to cooperation. While repeated 2-person prisoner's dilemmas can lead to cooperation under suitable conditions, this is not likely in n-person anonymous prisoner's dilemmas for two reasons: Firstly, in n-person settings it is not possible to punish specific agents with reciprocal defection. Secondly, if the setting is anonymous, it is not even possible to identify defectors, rendering punishment impossible. The model I propose enables cooperators to do well, because it introduces social structure. The key to successful cooperation is a clustering of cooperators, and an exclusion of defectors.

The problems of collective action and public goods provision have concerned political philosophers and economists for a long time. A lot of energy has been invested into explaining why, empirically, much more cooperation occurs than standard theory of collective action predicts. A focus on repeated games surely points in the right direction, but it only goes half the way. To understand the possibility of cooperation in n-person games with anonymous contributions, one has to add social structure. This move is particularly attractive because it makes social structure endogenous. Therefore, the model captures an important aspect of social interaction in reality: Social structure and the success of social interactions are in a dynamic relation with each other. Mutually beneficial interaction reinforces social relations, exploitation weakens it.

This phenomenon is well-known from real settings. For instance, when people venture into joint projects (founding a company, sharing a flat, writing a paper together, etc.), each participant can either contribute or free-ride. It is often difficult to detect free-riding, and even if it can be detected, it is difficult to punish the defector efficiently. Rather, people choose not to continue interaction in groups where the outcome is disappointing. Learning from experience, agents change the social structure of the environment by sticking with groups where free-riding is rare, and staying away from groups where free-riding is common. Individuals willing to cooperate try to cluster in groups with other cooperators, and try to exclude those who defect. What is remarkable about the simulation results is the success of this strategy, even if the information available to the agents is very limited. It is not necessary to track down specific defectors, it suffices to observe the collective outcome and change ties to other agents in response.

I have mentioned mundane examples of cooperation such as flatsharing, co-authoring papers, or running a company as a group of shareholders. However, Machiavelli and Hobbes remind us that the most fundamental problem of cooperation is about life and

death: The problem of providing security to live peacefully with each other. In a state of anarchy, where a central provision of policing and security is not possible, security becomes an n-person prisoner's dilemma. Hobbes reminds us that everyone can kill everyone in a (Hobbesian) state of nature:

> "NATURE hath made men so equal in the faculties of body and mind as that, though there be found one man sometimes manifestly stronger in body or of quicker mind than another, yet when all is reckoned together the difference between man and man is not so considerable as that one man can thereupon claim to himself any benefit to which another may not pretend as well as he. For as to the strength of body, the weakest has strength enough to kill the strongest, either by secret machination or by confederacy with others that are in the same danger with himself." (Hobbes, 1996, ch. 13)

When government fails, murderers are no longer kept in check by the threat of punishment. Everyone has to fight for themselves, and the public good of peace can no longer be provided. In these horrific situations, fleeing to safer areas is often the only option. The large-scale move of refugees in civil wars or failed states can be understood as the desperate attempt to find a safe haven of mutual cooperation in the most basic sense of cooperation: not killing each other. The model discussed in this chapter is certainly much too simple to be applied to such complex problems, but with some caveats one could draw the conclusion that overcoming a state of anarchy requires a formation of new local "clusters" of cooperation based on processes of inclusion and exclusion. The model could suggest that states of prolonged anarchy are likely to be followed by a phase of localisation, where villages or clans form cores of cooperation.

Clearly, much more methodological, theoretical, and empirical work is needed to apply computational models to such complex problems. Nevertheless, the example suggests that the potential areas of application for computational simulations are huge.

The search for peace is probably the most important example of a social dilemma. Again, this is a case where norm compliance is collectively beneficial. A social norm of cooperation could prescribe restraint to keep the peace. The problem is that compliance is difficult to enforce when other agents defect and are able to exploit those who exercise restraint. In these circumstances, social exclusion and the clustering of cooperators is a social process that can help to maintain high levels of cooperation. The subtle mechanism of network re-formation allows cooperators to move away from defectors. This indicates that the enforcement of social norms of cooperation is likely to be linked with group formation and exclusion processes.

Conclusion

8. Conclusion

I have said in the introduction that this thesis has a theory part and a modelling part. The theory part can be divided further into a moral theory component and a methods component. The moral theory component (chapter 3) offers a framework for realistic moral theory by using a possible worlds semantic. The methods component (chapters 4 and 5) discusses foundational issues regarding computational models and evolutionary arguments for the explanation of social norms. The modelling part (chapters 6 and 7) presents two models of exclusion mechanisms to maintain cooperation and to enforce social norms. Taken together, the two parts outline an argument for a new way to conduct a positive analysis of social norms and moral systems. The argument has run like this:

Compliance with social norms leads to a social dilemma because mutual compliance is in the collective interest, but individuals are often better off defecting. Therefore, social norms have to be enforcable, otherwise they are likely to collapse, given that those agents who comply with the norms are challenged by those who do not. Norms that can be enforced are feasible, those that cannot are utopian. Norm enforcement is a dynamic social process. To understand which social norms are feasible, we have to model the mechanisms of enforcement. Exclusion and group formation are possible enforcement mechanisms, and the dynamic simulations demonstrate that they can be effective. Thus, I conclude that mechanisms of exclusion play an important role in the enforcement of social norms.

I briefly rehearse the core arguments, discuss possible extensions, and offer an outlook for a future research programme towards the end of this conclusion.

8.1. Cooperation and Social Norms

Cooperation is a dilemma when individuals are tempted to defect, but collectively better off when everyone cooperates. The prisoner's dilemma embodies this problem for the two-person case, the public goods dilemma (or n-person prisoner's dilemma) for the multi-person case. Most real life problems involve more than two people. Therefore,

the multi-person case has received much more attention in this thesis. Unfortunately, cooperation in multi-person prisoner's dilemmas is an even harder problem than in the two-person case. For the two-person case, there is good hope that cooperation can emerge when games are repeated (infinitely, or with a low probability of ending the game for every round). This insight, which has received much popular attention with Axelrod's (1984) computer tournaments, does not extend easily to n-person games: While reciprocal strategies (like TIT-FOR-TAT) work well in the two-person case, because reciprocal behaviour directly targets the defector, no such direct targetting is possible in the n-person case. Therefore, it is impossible to "punish" only defectors, without also hurting other cooperators. Things get even worse when players cannot observe what others do. In these anonymous n-person prisoner's dilemmas, cooperation is almost impossible to maintain.

According to the Folk theorem, mutual cooperation can nevertheless be a Nash equilibrium in the n-person case. This can be demonstrated by imagining that agents use "trigger strategies". If agents cooperate as long as all other agents cooperate, and defect for all coming rounds as soon as one other agent defects, and if it is common knowledge that agents follow this strategy, then defection does not pay because it destroys all future cooperation gains. Therefore, mutual cooperation is a Nash equilibrium in the infinite n-person prisoner's dilemma. However, it is not a stable equilibrium under the reasonable assumption that agents make occasional mistakes. Either mistakes lead to an immediate breakdown of cooperation, or agents give up the radical "trigger strategy" in order to cope with mistakes, which renders the threat of the trigger strategy incredible. The upshot is that cooperation is difficult, if not impossible, to maintain in repeated n-person prisoner's dilemmas and public goods games, unless other forms of sanctioning or incentives are introduced.

A *social norm of cooperation* is a norm that prescribes actions, which, when obeyed by all participating parties, produces greater average payoffs, compared to outcomes where the norm is not in place and agents pursue individual payoff maximisation. Therefore, a social norm of cooperation prescribes cooperation in a social dilemma to secure a mutual cooperation gain. However, compliance with the social norm of cooperation is itself a social dilemma. Whether that dilemma can be solved depends on how well the norm is enforceable. The dilemma posed by norm compliance and its solutions are therefore crucial for understanding how groups and societies cooperate with each other and keep defectors in check.

Social norms of cooperation must be distinguished from conventions. A convention is a norm that coordinates actions, but if everyone knows which convention is valid there

is no necessity to enforce the convention. Therefore, conventions have not been under the remit of my analysis.

8.2. Realistic Moral Theory

By *modus tollens*, it follows from the "ought implies can principle" (OICP) that there is no "ought" for unfeasible actions. If an action cannot be performed, it cannot be obligatory to perform it (exceptions occur when an agent is responsible for being unable to perform an action). This is relevant for both positive and normative analyses of norms: We need to know which actions are feasible under which conditions to understand why we have the norms we have, but also to adjust our normative claims about obligations and responsibility.

By using a possible worlds semantic I have been able to distinguish between the set of permissible worlds, the set of feasible worlds, and the actual world. The set of permissible worlds is the set of all worlds where all relevant norms are obeyed by all agents. When the norms are demanding, this set is small, when the norms are less demanding, it is larger. The set of feasible worlds is the set of all accessible worlds, given the actual world. Different concepts of feasibility lead to different sets of feasible worlds.

I have discussed four concepts of feasibility: Physical feasibility, physical feasibility with a *ceteris paribus* assumption, psychological feasibility, and social feasibility. I have shown that the physical feasibility concepts are not adequate for defining feasibility for the present purposes: What is feasible for real human agents depends not only on what is physically feasible, but also on the agents' psychological and social limitations. The psychological feasibility concept assumes that agents are not saints and therefore limited in their actions by their psychological conditions. However, I have argued that the psychological feasibility concept does not completely capture what is typically feasible for socially embedded agents. What is feasible for such an agent depends not only on the agent him- or herself, but also on the actions of the others. The feasibility of actions changes dynamically, depending on how the actions change the social environment agents are operating in. Therefore, which actions are feasible for an agent must be determined by conducting a dynamic analysis of social interactions. This is what the agent-based models I have developed in chapters 6 and 7 do.

Once the concept of feasibility is suitably defined—and the precise definition can vary from case to case—it is possible to distinguish between realistic and utopian norms. If the intersection of the set of permissible worlds and the set of feasible worlds is empty,

then the norms that determine the set of permissible worlds are utopian. But if the intersection is non-empty, the norms are realistic in the sense that a permissible world is accessible for agents, given physical, psychological, and social constraints.

Dynamic feasibility analyses allow us to move towards realistic moral theory. The notion of "realistic moral theory" is related to Rawls's "non-ideal theory". Staking out the limits between utopian and realistic moral theory has important implications for normative debates. For example, the question of moral responsibility depends on whether agents can realistically be expected to meet their obligations. In addition, the distinction between utopian and realistic moral theory is also relevant for a positive analysis of moral systems. It helps to explain why certain norms have emerged and stabilised, while others have not.

The analysis of feasibility constraints for moral theory also allows us to build a bridge between evolutionary game theory and moral theory. Following Skyrms's idea that evolutionary game theory could provide "some interesting contraints" for moral theory, I have argued that an evolutionary analysis of norms can help us to understand which norms are evolutionarily feasible and which are not. This does not imply that evolutionary game theory can predict which moral norms emerge, but at least it helps us to understand the feasibility constraints for realistic moral theory.

8.3. Agent-Based Models for the Social Sciences

In this thesis I have taken a positive, model-based approach towards issues in the social sciences and philosophy. A model must be a representation of its target system. I have assumed with many philosophers of science that a model is a representation of its target system in virtue of a similarity relation between model and target system. However, the similarity relation is difficult to define, especially for models in the social sciences. Often we do not understand the target systems well. In these cases, it is usually preferable to work with simple, highly idealised models (HIM). HIM should focus on specific slices of reality. More precisely, social scientists should aim to isolate social mechanisms. A social mechanism singles out a causal relation between input and output variables. Social mechanisms are integrated in a more complex system of interacting processes in the target system. Therefore, one usually cannot observe an isolated social mechanism in reality. Modelling, however, allows the researcher to explore a hypothesised social mechanism under rigorous control conditions.

To model social mechanisms and abstract from other intervening processes, the social scientist has to use two distinct processes of idealisation: isolation and generalisation.

8. Conclusion

Isolation is necessary for singling out the social mechanism and omitting all other factors that are not directly relevant. Generalisation is necessary for demonstrating that the mechanism is robust in the sense that it applies to many different target systems, and not only to very specific constellations. Generality is particularly important when the target system is not well understood. To argue that a general mechanism is active, it often suffices to have a coarse-grained understanding of the target system, while arguing for the presence of a less general mechanism requires a more fine-grained understanding.

Ideally, the model assumptions should be supported by empirical evidence, or the model outcomes should be tested against empirical data. But when models are highly idealised and the target system complex, this is not always possible.

Agent-based models are useful for analysing social mechanisms because they enable the researcher to explore dynamic, complex systems, where a purely analytical approach fails. Complex systems can arise even when the hypothesised mechanism is simple. In the analyses of cooperation on networks, for example, I have operated with extremely simple assumptions about agent behaviour and social space. Nevertheless, the ensuing dynamics are complex. In these cases, computational agent-based models can help us to understand the properties of the system by running many simulations with different parameter values.

In chapter 5 I have discussed possible applications of evolutionary theory to gain a better understanding of social processes. In my view, one should not expect too much from evolutionary explanations of contemporary social processes. From the perspective of biological evolution, humans are adapted to an environment that is radically different from modern societies. Biological evolution can offer us explanations for some fundamental psychological dispositions, but is unlikely to provide an explanatory key for behaviour that is influenced by social norms. Cultural evolution might offer some explanatory leverage for social norms and cultural phenomena, but these explanations are weak at best if there is no clear concept of fitness for cultural replicators such as "memes". The problem is that many proponents of cultural evolution employ different concepts of fitness. If one understands fitness for a cultural replicator (a "meme" or something equivalent) as the replicator's factual success to replicate, then this reduces the theory to the tautology "the fitness of a replicator is the fitness of a replicator". If, however, one uses a more substantial definition, for example linking the fitness of a replicator to the utility of its host, then the theory is stronger, but also prone to be refuted by empirical evidence.[1]

[1] Lina Eriksson (personal communication) once remarked that researchers often endorse a very weak definition of fitness at the conceptual level, but then slip to a stronger concept in their argument.

8.4. Models of Exclusion and Cooperation

In this thesis I have focused on mechanisms of group formation and exclusion to understand how the compliance problem, and the more general problem of cooperation in public goods games, are solved. In the first model, agents have some rather unreliable but better than random *ex ante* information about the disposition of other agents to play a certain strategy. Cooperators can be successful and avoid defectors if they are able to pool their individual information about the disposition of other agents, and if they can decide on the admission of agents into their group. In these cases, cooperators do well because they cluster with other cooperators, and avoid exploitation by defectors. The model looks abstract, but it captures one key element of human sociality neatly: To secure cooperation, humans form groups of likeminded people, and they use an *ex ante* selection of people to make sure that cooperation works. This may be one reason why people are not only interested in the actions of others, but also in the motives for their actions. The interest in motives reveals the interest in finding out about cooperative dispositions. Is someone cooperating just because he is observed, or would he still cooperate even if there was an opportunity for free-riding? Answering this question is important if the *ex ante* selection of interaction partners plays a role in the maintenance of cooperation.

The second model also focuses on n-person prisoner's dilemmas, but in this model agents react to previous game outcomes and change the social space of interaction. Agents are represented as vertices of a network. In each round, each agent plays a public goods game with all other connected agents. The contributions to these games are anonymous. This means that agents can usually not infer who has cooperated and who has defected in the game. However, they do learn the outcomes of the games they participate in and can infer how many cooperators and defectors are connected to them. The agents can use this information to influence the structure of the network. After playing, each agent in the network can choose to delete one edge to one of her neighbours. Deleted edges are replaced with random edges in the network. The opportunity to delete edges gives the agents a chance to disconnect from defectors. This in turn can lead to a separation in the network, where cooperators are only connected with cooperators, and defectors only with defectors. My simulations have shown that this separation occurs robustly for a wide range of parameter settings.

The simulation of agents on networks is important for understanding the emergence of social cooperation because it models an important aspect of social interaction: Peo-

I agree.

ple can often decide with whom they want to interact, and with whom interaction is not worthwhile. They build and rebuild their network of contacts and acquaintances to make sure that their interactions go well. The existence of a dynamic social structure explains why dilemmas of cooperation and norm compliance can often be solved. Defection is "punished" with a loss of contacts and, ultimately, exclusion from the network of cooperators.

8.5. Policy Implications

The last three sections rehearsed the core arguments of the thesis. I now consider some possible policy implications.

In the two models in chapters 6 and 7 I have analysed group formation and exclusion as special forms of norm enforcement. In these models, cooperators aim to exclude defectors from anonymous n-person prisoner's dilemmas. Exclusion works when the game can be "localised" in the sense that some agents can be excluded from participation, while the others still play. If the n-person prisoner's dilemma is understood as a public goods game, one can say that the "localisation" converts the public good into a club good (see Buchanan, 1965). For example, team work with changing team composition allows agents to choose their playing partners, and avoid others. The team product is a club good. Even if it is not possible to identify directly who is "shirking" or free-riding in the team, it might be possible for cooperators to exclude defectors, either by pooling weak *ex ante* information, or by using the aggregate outcomes *ex post* as cues to move away from defectors, i. e. to exit the club (compare Hirschman, 1970). I have predicted that these processes are important for the formation and regrouping of business teams, or of research groups and collaboration networks in the academic world.

It is also plausible that *ex post* mechanisms of selection are taking place when people select neighbourhoods to live in: The quality of life in a neighbourhood depends on how other neighbours behave, and this gives rise to public goods dilemmas. For instance, it is collectively preferable if agents carry their broken appliances to the next recycling centre—but it is often more convenient to dump them in the street. Identifying and prosecuting perpetrators is often difficult or impossible. Instead, people "vote with their feet" and try to leave run-down quarters if they can. The increasing number of gated and managed developments in many cities might be driven by this mechanism.

I have argued that processes of group formation and exclusion are important mechanisms to enforce norms locally. Finding the right cooperation partners *ex ante*, or moving away *ex post* is often a cheap form of enforcement. However, the models dis-

cussed here are not suitable when exclusion is not possible. For example, the reduction of greenhouse gases (such as CO_2 and methane) is a public goods dilemma. But it is not possible to exclude agents from this "game". Therefore, exclusion and group formation do not work as sanctioning mechanisms. But exclusion might still work when different games are linked. For instance, it may be possible to enforce a reduction of emissions by threatening to exclude an agent in other cooperative projects. Companies dealing with environmentally aware customers may feel this form of pressure already: A company with a bad emissions record may suffer from an image-loss and ultimately from lower revenues when customers turn to other companies. With issue-linkage, the sanctioning power of exclusion increases.

The network model in chapter 7 also alludes to one of the most fundamental problems of political philosophy: the issue of security. If one follows Hobbes's view that the provision of security poses a social dilemma, then the network model suggests a tentative idea how social structure may help to solve the dilemma. Hobbes suggests that there are two types of agents: those who "would be glad to be at ease within modest bounds" and those who are "taking pleasure in contemplating their own power in the acts of conquest, which they pursue farther than their security requires" (Hobbes, 1996, ch. XIII). The first type of agent resembles a cooperator, the second a defector in the dilemma. For Hobbes, the only viable solution out of the dilemma is the "Leviathan". Only the authorisation of one individual to exercise absolute power could keep the defectors in check. However, the network model suggests that cooperators do not necessarily need a Leviathan. If they are able to change the social structure to their advantage, they can avoid defectors and exclude them from their social interactions. More practically speaking, peace-seeking individuals try to avoid more aggressive ones. Even though peace seekers are too weak to punish their more aggressive counterparts for their "defection" directly, they can deny them cooperation gains, and they can invest their own cooperation gains into a collective defence. This may be a way to solve the dilemma without assuming a Leviathan.

This raises some thoughts regarding modern challenges posed by failed states and societies in civils wars or anarchy. In such broken societies the enforcement of norms, both social and legal, has broken down. Once this has happened, the usual means of state enforcement fail. It is now extremely hard to reinstate law and order because the social norms that had previously supported the legal enforcement of norms have now collapsed. A normal police force is usually not able to reinstate order. In a working state, the official policing was complemented by "social policing". Once that "social policing" is gone, the official policing is insufficient to regain control. In these disastrous

situations, we often see large refugee movements. In the spirit of the network model, one can interpret this as a desperate attempt to form new clusters of cooperation.

8.6. New Questions

I have argued that computational agent-based models are the adequate tool for understanding feasibility constraints for social norms from a social, dynamic perspective. My two models focus on group formation and exclusion as two special mechanisms of subtle enforcement. While I think that these mechanisms play an important role in many situations, they are not the only subtle sanctioning mechanisms. Different situations require different mechanisms. In many situations, indirect reciprocity and reputation are crucial. Indirect reciprocity is particularly powerful when agents can be observed and when people keep track of each other's reputation. Other dynamics arise when people cannot reciprocate, but have the opportunity to punish (Henrich and Boyd, 2001). In addition, more complicated dynamics are at play when the agents link different games. In reality, people often reciprocate (or indeed retaliate) in a different domain. For instance, if I think that my neighbour creates irresponsible environmental damage by buying a Hummer SUV, it is not in my interest to "reciprocate" by buying one too. Rather, I might use different sanctioning mechanisms, such as not inviting him to my parties or bad-mouthing him in the neighbourhood. As already observed, linking issues gives opportunites of reciprocation or punishment that are otherwise unavailable.

Often the "goods" involved in these games of reciprocation, punishment, or exlusion are not even of a material nature. Brennan and Pettit (2004) point out that many people desire to be held in *esteem* by other people, and that withdrawing esteem is a powerful social sanctioning mechanism. Esteem is also directly linked with social norms because social norms define which actions deserve esteem or disesteem. Empirical evidence suggests that the desire for esteem or social approval is indeed an important factor for humans in cooperation dilemmas (Gächter and Fehr, 1999). More models of subtle norm enforcement are needed to understand these mechanisms.

It would be interesting to explore the network model with more variations. One option is to introduce a cost to the deleting of an edge.[2] Also, it would be worthwhile to know how the model behaves if agents adopt mixed strategies, either for their game strategies or their network strategies. I am thinking of "smarter" defectors, who are able to distinguish between situations where they are easily spotted as defectors, and situations where free-riding is less risky. The defectors could adapt their game strategy

[2] I am grateful to Ben Kerr (personal communication) who suggested this idea to me.

accordingly and only defect when they are likely to get away with it. Similarly, cooperators could also refine their strategies by conditionalising their cooperation. Finally, an introduction of a learning mechanism or an evolutionary dynamic can be introduced.

I have argued that the assumptions of my models are empirically supported. However, for a more rigorous empirical support it would be nice to simulate the mechanisms described in the models in a lab experiment. I think that the translucency model should be relatively easy to implement because it requires only one round of playing after the group formation process. A direct test of the network model is more difficult because many rounds are needed before cooperators and free riders are likely to separate.

8.7. Positive, Analytical Theories of Social Norms

The overarching argument of this thesis suggests to look at social norms from a positive, model-based perspective. I have said that the overarching argument is an outline or a sketch. While I believe that the arguments presented in each chapter are valid and their premises well supported, I do not make an equally strong claim regarding the overarching argument. Much more theoretical and empirical work is needed to develop this sketch into a theory proper. In this last section I want to make some proposals on how the sketch I offer can be turned into a research programme towards a positive, analytical theory of social norms. These last remarks are necessarily and deliberatively speculative.

There has been an increasing interest in the positive analysis of social and in particular moral norms in recent years. I have mentioned Brian Skyrms's (1996) idea to derive constraints for moral theory from evolutionary game theory. Binmore (2004, 1998, 1994) offers a more radical attempt to use evolutionary game theory with the aim to naturalise moral theory. Ruse (1995) also tries to develop a naturalistic theory of morality based on evolutionary theory. I think that Skyrms's line is most promising, and my considerations regarding realistic moral theory are meant to explicate and extend his approach.

In addition to Skyrms's careful theoretical reasoning, I also anticipate an increasing influence of empirical work in the positive analysis of social norms and moral theory. Several disciplines are relevant in that regard. Behavioural economists and experimental social psychologists observe choices in dilemma situations. These dilemma situations can be controlled rigorously by the researcher in the lab. The results accumulated in recent years (for a review see Camerer, 2003) shed new light on the behaviour of humans in dilemmas of cooperation and questions regarding social norms (Fehr and Fischbacher,

2004). Fischbacher and Gächter (2006) find that there are different types of agents in public goods games: They identify around 50% "conditional cooperators" and around 25% free riders in their experiments (the other agents are more difficult to classify). Further research in this area is likely to uncover the actual behaviour of humans in dilemma situations in great detail. The results from behavioural economics can also be connected to evolutionary theory (Bowles and Gintis, 2003; Fehr and Henrich, 2003).

Another promising route of empirical research into the nature of norms is moral psychology (see for example Greene, in press; Haidt, 2001). This research is driven by experiments and techniques of neuro-imaging. While there are still many question marks about the results and their interpretation, this line of enquiry is likely to change our understanding of normative reasoning. New approaches in "experimental philosophy" also link up with moral psychology (Knobe, 2003).

I predict that computational models will be one of the hinges between theoretical reasoning and empirical research. For the time being, it is unlikely that a complete theory of human normative reasoning can be developed. It is more likely that we have to deal with partial results. We may soon be able to understand some mechanisms within a much more complex process, but we are far from able to understand all mechanisms and their relations to each other. In these situations, computational simulations can help us to explore the implications of different mechanisms and their dynamic interactions. For instance, agent-based models can be refined by using assumptions based on results from behavioural economics and moral psychology. These models could then give us the opportunity to enhance our understanding of dynamic social interactions. The results could in turn feed back into further theoretical and empirical research.

Whether the new, positive approaches to normative reasoning will ever replace normative moral theory by an entirely naturalistic theory is difficult to predict. Proponents of moral non-naturalism deny that a reduction of moral properties to non-moral properties is possible (Ridge, 2006). I do not espouse a meta-ethical position in this thesis, but I share the intuition that a reductionist form of moral naturalism is unsatisfying. Natural properties alone seem to lack the justificatory and motivational force of moral properties. Even if we understood human morality perfectly well in positive terms, it would not give us reasons or motivation to act morally. A reductionist naturalistic theory of morality does not capture the importance of normativity for human thought and action. Having said this, there are of course many subtly different naturalistic meta-ethical stances one can take, which I cannot discuss here. All I want to claim is that for the time being, it is unlikely that an encompassing positive, naturalistic theory emerges, partly because of practical limitations, partly because it seems theoretically

implausible. Therefore, it is reasonable to view research into positive aspects of moral theory not as competitor to normative theory, but rather as a useful complement. As long as we are without an encompassing theory—and this might be for a long time—we can at least use the knowledge we have so far to improve our understanding of the constraints on human social norms.

In this thesis I have discussed the different components of such a research programme. I have argued that positive theory and normative theory are connected through the concept of feasibility. I have discussed methodological and philosophy-of-science questions regarding the use of computational models in the social sciences. Finally, I have given two examples of how agent-based models can be utilised to work towards a better understanding of social norms and moral theory, particularly with regard to norm compliance. The results I have presented here are only the first step in a rapidly developing field of research. More work is needed.

Bibliography

Akerlof, George A. 1970. "Market For Lemons – Quality Uncertainty And Market Mechanism." *Quarterly Journal Of Economics* 84(3):488–500.

Akerlof, George A. and Rachel E. Kranton. 2000. "Economics and Identity." *Quarterly Journal of Economics* 115(3):715–754.

Alexander, Jason McKenzie. 2003*a*. Evolutionary Game Theory. In *Stanford Encylopedia of Philosophy (Summer 2003 edition)*, ed. Edward N. Zalta.
URL: *http://plato.stanford.edu/archives/sum2003/entries/game-evolutionary/*

Alexander, Jason McKenzie. 2003*b*. "Random Boolean Networks and Evolutionary Game Theory." *Philosophy of Science* 70:1289–1304.

Alexander, Jason McKenzie. 2007. *The Structural Evolution of Morality.* Cambridge et al.: Cambridge University Press.

Aumann, Robert J. 1989. *Lectures on game theory.* Boulder: Westview Press.

Aunger, Robert, ed. 2000. *Darwinizing Culture: The Status of Memetics as a Science.* Oxford and New York: Oxford University Press.

Austen-Smith, David and Jeffrey S. Banks. 1996. "Information Aggregation, Rationality, and the Condorcet Jury Theorem." *American Political Science Review* 90(1):34–45.

Axelrod, Robert M. 1984. *The evolution of cooperation.* New York: Basic Books.

Axtell, Robert L., Joshua M. Epstein, Jeffrey S. Dean, George J. Gumerman, Alan C. Swedlund, Jason Harburger, Shubha Chakravarty, Ross Hammond, Jon Parker and Miles Parker. 2006. Population Growth and Collapse in a Multiagent Model of the Kayenta Anasazi in Long House Valley. In *Generative Social Science: Studies in Agent-Based Computational Modeling.* Princeton and Oxford: Princeton University Press pp. 117–129.

Bacharach, Michael. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory.* Princeton and Oxford: Princeton University Press.

Batterman, Robert W. 2002. "Asymptotics and the Role of Minimal Models." *British Journal for the Philosophy of Science* 53:21–38.

Bernhard, Helen, Ernst Fehr and Urs Fischbacher. 2006. "Group Affiliation and Altruistic Norm Enforcement." *American Economic Review* 92(2):217–221.

Bicchieri, Christina. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms.* Cambridge et al.: Cambridge University Press.

Binmore, Ken. 1992. *Fun and games: a text on game theory.* Lexington, Mass: D.C. Heath.

Binmore, Ken. 1994. *Playing Fair: Game Theory and the Social Contract I.* Cambridge, Mass: MIT Press.

Binmore, Ken. 1998. *Just Playing: Game Theory and the Social Contract II.* Cambridge, Mass: MIT Press.

Binmore, Ken. 2001. "Evolutionary Social Theory: Reply to Robert Sugden." *The Economic Journal* 111(469):244–248.

Binmore, Ken. 2004. "Reciprocity and the Social Contract." *Politics, Philosophy & Economics* 3(1):5–35.

Binmore, Ken. 2005. *Natural Justice.* New York: Oxford University Press.

Binmore, Ken. 2006. "Why do people cooperate?" *Politics, Philosophy & Economics* 5(1):81–96.

Binmore, Ken. 2007. *Playing for real: a text on game theory.* New York: Oxford University Press.

Blackmore, Susan. 1999. *The meme machine.* Oxford: Oxford University Press.

Bowles, Samuel and Herbert Gintis. 2003. Origins of Human Cooperation. In *Genetic and Cultural Evolution of Cooperation*, ed. Peter Hammerstein. Cambridge (MA) and London: M.I.T. Press pp. 429–443.

Boyd, Robert and Peter J. Richerson. 1985. *Culture and the Evolutionary Process.* Chicago and London: University of Chicago Press.

Bibliography

Boyd, Robert and Peter J. Richerson. 2005. *The Origin and Evolution of Cultures.* Oxford and New York: Oxford University Press.

Brandt, Hannelore and Karl Sigmund. 2004. "The logic of reprobation: assessment and action rules for indirect reciprocation." *Journal of Theoretical Biology* 231(4):475–486.

Brennan, Geoffrey. 1996. Selection and the Currency of Reward. In *Theory of Insititutional Design*, ed. Robert E. Goodin. Cambridge University Press.

Brennan, Geoffrey and Philip Pettit. 2000. "The Hidden Economy of Esteem." *Economics and Philosophy* 16:77–98.

Brennan, Geoffrey and Philip Pettit. 2004. *The Economy of Esteem: An Essay on Civil and Political Society.* Oxford: Oxford University Press.

Brosig, Jeanette. 2002. "Identifying cooperative behavior: some experimental results in a prisoner's dilemma game." *Journal of Economic Behavior & Organization* 47(3):275–290.

Buchanan, James M. 1965. "An Economic Theory of Clubs." *Economica* 32(125):1–14.

Camerer, Colin. 2003. *Behavioral game theory: experiments in strategic interaction.* New York, N.Y: Russell Sage Foundation.

Camerer, Colin F. and Ernst Fehr. 2006. "When does "economic man" dominate social behavior?" *Science* 311(5757):47–52.

Cartwright, Nancy. 1999. *The Dappled World: A Study of the Boundaries of Science.* Cambridge: Cambridge University Press.

Cinyabuguma, Matthias, Talbot Page and Louis Putterman. 2005. "Cooperation under the threat of expulsion in a public goods experiment." *Journal of Public Economics* 89(8):1421–1435.

Congleton, Roger D. and Viktor J. Vanberg. 2001. "Help, harm or avoid? On the personal advantage of dispositions to cooperate and punish in multilateral PD games with exit." *Journal of Economic Behavior & Organization* 44(2):145–167.

Cowen, Tyler. 2007. "The Importance of Defining the Feasible Set." *Economics and Philosophy* 23(01):1–14.

Danielson, Peter A. 1992. *Artificial Morality. Virtuous Robots for Virtual Games.* London and New York: Routledge.

Dawes, Robyn M., Alphons J. C. van de Kragt and John M. Orbell. 1988. "Not Me or Thee but We: The Importance of Group Identity in eliciting Cooperation in Dilemma Situations: Experimental Manipulations." *Acta Psychologica* 68:83–97.

Dawkins, Richard. 1989. *The Selfish Gene.* New ed. Oxford and New York: Oxford University Press.

Dawkins, Richard. 1994. "Burying the Vehicle." *Behavioral and Brain Sciences* 17(4):616–617.

de Marchi, Scott. 2005. *Computational and Mathematical Modeling in the Social Sciences.* Cambridge et al.: Cambridge University Press.

de Vos, Henk, Rita Smaniotta and Donald A. Elsas. 2001. "Reciprocal Altruism under Conditions of Partner Selection." *Rationality and Society* 13(2):139–183.

Dean, Jeffrey S., George J. Gumerman, Joshua M. Epstein, Robert L. Axtell, Alan C. Swedlund, Miles T. Parker and Stephen McCarroll. 2000. Understanding Anasazi culture change through agent-based modeling. In *Dynamics in Human and Primate Societies: Agent-Based Modeling of Social and Spatial Processes*, ed. T. A. Kohler and G. J. Gumerman. Oxford University Press pp. 179–205.

Dennett, Daniel C. 1996. *Darwin's Dangerous Idea: Evolution and the Meanings of Life.* London: Penguin.

Dennett, Daniel C. 2006. *Breaking the Spell: Religion as a Natural Phenomenon.* London: Penguin.

Dietrich, Franz and Christian List. 2004. "A model of jury decisions where all jurors have the same evidence." *Synthese* 142:175–202.

Ellickson, Robert C. 1998. "Law and Economics Discovers Social Norms." *Journal of Legal Studies* 27(2):537–552.

Epstein, Joshua M. 2006. *Generative Social Science.* Princeton University Press.

Epstein, Joshua M. and Robert Axtell. 1996. *Growing artificial societies: social science from the bottom up.* Washington: Brookings Institution Press.

Estlund, David, Waldron. Jeremy, Bernard Grofman and Scott L. Feld. 1989. "Democratic Theory and the Public Interest: Condorcet and Rousseau Revisited." *American Political Science Review* 83(4):1317–1340.

Fagiolo, Giorgio, Marco Valente and Nicolaas J. Vriend. 2005. Segregation in Networks. Working Papers 549 Queen Mary, University of London, Department of Economics. **URL:** *http://ideas.repec.org/p/qmw/qmwecw/wp549.html*

Fehr, Ernst and Joseph Henrich. 2003. Is strong Reciprocity a Maladaptation? On the Evolutionary Foundations of Human Altruism. In *Genetic and Cultural Evolution of Cooperation*, ed. Peter Hammerstein. Cambridge, Mass. and London: MIT Press pp. 55–82.

Fehr, Ernst and Urs Fischbacher. 2004. "Social norms and human cooperation." *Trends in Cognitive Sciences* 8(4):185–190.

Fischbacher, Urs and Simon Gächter. 2006. "Heterogeneous social preferences and the dynamics of free riding in public goods." *CeDEx Discussion Paper* 1.

Frank, Robert H. 1987. "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" *The American Economic Review* 77(4):593–604.

Frank, Robert H. 1988. *Passions within Reason.* New York and London: W. W. Norton & Company.

Frank, Robert H., Thomas Gilovich and Dennis T. Regan. 1993. "The Evolution of One-Shot Cooperation – an Experiment." *Ethology and Sociobiology* 14(4):247–256.

Frigg, Roman. 2006. "Scientific Representation and the Semantic View of Theories." *Theoria* 55:49–65.

Fu, Feng, Xiaojie Chen, Lianghuan Liu and Long Wang. 2007. "Promotion of cooperation induced by the interplay between structure and game dynamics.". **URL:** *http://arXiv.org/pdf/physics/0701322*

Fudenberg, Drew and Eric Maskin. 1990. "Evolution and Cooperation in Noisy Repeated Games." *The American Economic Review* 80(2):274–279.

Gauthier, David P. 1986. *Morals by Agreement.* Oxford: Clarendon.

Gächter, Simon. 2006. "Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications." *CeDEx Discussion Paper* 03.

Gächter, Simon and Christian Thöni. 2005. "Social Learning and Voluntary Cooperation among Like-minded People." *Journal of the European Economic Association* 3:303–314.

Gächter, Simon and Ernst Fehr. 1999. "Collective action as a social exchange." *Journal of Economic Behavior & Organization* 39:341–369.

Giere, Ronald N. 1988. *Explaining Science: A Cognitive Approach.* Chicago: University of Chicago Press.

Gigerenzer, Gerd. 2001. The Adaptive Toolbox. In *Bounded Rationality: The Adaptive Toolbox*, ed. Gerd Gigerenzer and Reinhard Selten. Cambridge (MA) and London: MIT Press pp. 37–50.

Gil-White, Francisco J. 2005. Common Misunderstandings of Memes (and Genes): The Promise and the Limits of the Genetic Analogy to Cultural Transmission Processes. In *Perspectives on Imitation: From Neuroscience to the Social Sciences*, ed. Susan Hurley and Nick Chater. Vol. 2 London: MIT Press pp. 317–338.

Gilbert, Nigel and Klaus G. Troitzsch. 2005. *Simulation for the Social Scientist.* 2nd ed. Open University Press.

Gintis, Herbert. 2000*a*. *Game Theory Evolving.* Princeton: Princeton University Press.

Gintis, Herbert. 2000*b*. "Strong Reciprocity and Human Sociality." *Journal of Theoretical Biology* 206:169–179.

Gintis, Herbert. 2007. "A framework for the unification of the behavioral sciences." *Behavioral and Brain Sciences* 30:1–61.

Glennan, Stuart. 2002. "Rethinking Mechanistic Explanation." *Philosophy of Science* 69:S342–S353.

Godfrey-Smith, Peter. 2006. "The strategy of model-based science." *Biology and Philosophy* 21:725–740.

Goette, Lorenz, David Huffman and Stephan Meier. 2006. "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups." *American Economic Review* 96(2):212–216.

Greene, Joshua D. in press. The secret joke of Kant's soul. In *Moral Psychology*, ed. Walter Sinnott-Armstrong. Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development Cambridge, MA.: MIT Press.

Grüne-Yanoff, Till. 2006. "Agent-Based Simulation, Generative Science, and its Explanatory Claims.".
URL: *http://philsci-archive.pitt.edu/archive/00002784/01/ACEpaper060609.pdf*

Güth, Werner and Hartmut Kliemt. 2000. "Evolutionarily stable co-operative commitments." *Theory and Decision* 49(3):197–221.

Haidt, Jonathan. 2001. "The emotional dog and its rational tail: A social intuitionist approach to moral judgment." *Psychological Review* 108:814–834.

Hardin, Russell. 1982. *Collective action.* Baltimore: Johns Hopkins University Press.

Hausman, Daniel M. 1992. *The Inexact and separate science of economics.* Cambridge: Cambridge University Press.

Hausman, Daniel M. 2005. "Sympathy, Commitment, and Preference." *Economics and Philosophy* 21(1):33–50.

Hedström, Peter and Richard Swedberg. 1998. Social mechanisms: An introductory essay. In *Social mechanisms: an analytical approach to social theory*, ed. Peter Hedström and Richard Swedberg. Cambridge: Cambridge University Press pp. 1–31.

Henrich, Joseph and Richard McElreath. 2003. "The evolution of cultural evolution." *Evolutionary Anthropology: Issues, News, and Reviews* 12(3):123–135.

Henrich, Joseph and Robert Boyd. 2001. "Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas." *Journal of Theoretical Biology* 208:79–89.

Hirschman, Albert O. 1970. *Exit, voice, and loyalty: responses to decline in firms, organizations, and states.* Cambridge, Mass; London: Harvard University Press.

Hirshleifer, David and Eric Rasmusen. 1989. "Cooperation in a repeated prisoners' dilemma with ostracism." *Journal of Economic Behavior & Organization* 12(1):87–106.

Hobbes, Thomas. 1996. *Leviathan.* Rev. ed. Cambridge: Cambridge University Press. First published 1651.

James, Henry. 1986. *Daisy Miller.* London: Penguin. First published 1878.

Jones, Martin R. 2005. Idealization and Abstraction: A Framework. In *Idealization XII: Correcting the Model. Idealization and Abstraction in the Sciences*, ed. Nancy Cartwright and Martin R. Jones. Vol. 86 of *Poznan Studies in the Philosophy of the Sciences and Humanities* Amsterdam and New York: Rodopi pp. 173–217.

Kitcher, Philip. 1985. *Vaulting ambition: sociobiology and the quest for human nature.* Cambridge, Mass.; London: MIT Press.

Kliemt, Hartmut. 1996. Simulation and Rational Practice. In *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View*, ed. Rainer Hegselmann, Ulrich Mueller and Klaus G. Troitzsch. Dordrecht et al.: Kluwer.

Knobe, Joshua. 2003. "Intentional Action and Side Effects in Ordinary Language." *Analysis* 63:190–193.

Kohler, Timothy A. 2000. Putting Social Sciences Together Again: An Introduction to the Volume. In *Dynamics in Human and Primate Societies.* New York and Oxford: Oxford University Press.

Küppers, Günter and Johannes Lenhard. 2005. "Validation of Simulation: Patterns in the Social and Natural Sciences." *Journal of Artificial Societies and Social Simulation* 8(4).
URL: *http://jasss.soc.surrey.ac.uk/8/4/3.html*

Kurzweil, Ray. 2006. "Deep Fritz Draws: Are Humans Getting Smarter, or Are Computers Getting Stupider?".
URL: *http://www.kurzweilai.net/meme/frame.html?main=/articles/art0527.html*

Ledyard, John O. 1994. "Public Goods: A Survey of Experimental Research.". also published in: The handbook of experimental economics, John H. Kagel and Alvin E. Roth (eds.), Princeton (NJ): Princeton University Press, 1995.
URL: *http://ideas.repec.org/p/wpa/wuwppe/9405003.html*

Leimar, Olof and Peter Hammerstein. 2001. "Evolution of cooperation through indirect reciprocity." *Proceedings of the Royal Society London B* 268:745–753.

Lewis, David. 1969. *Convention: A Philosophical Study.* Cambridge et al.: Cambridge University Press.

Lewis, David. 1979. "Counterfactual Dependence and Time's Arrow." *Nous* 13(4):455–476.

Lewontin, Richard. C. 1970. "The Units of Selection." *Annual Review of Ecology and Systematics* 1:1–14.

Lieberman, Erez, Christoph Hauert and Martin A. Nowak. 2005. "Evolutionary dynamics on graphs." *Nature* 433(7023):312–316.

Lillehammer, H. 2003. "Debunking morality: evolutionary naturalism and moral error theory." *Biology & Philosophy* 18(4):567–581.

List, Christian. 2006. "Republican Freedom and the Rule of Law." *Politics, Philosophy & Economics* 5(2):201–220.

List, Christian and Robert E. Goodin. 2001. "Epistemic Democracy: Generalizing the Condorcet Jury Theorem." *Journal of Political Philosophy* 9(3):277–306.

Lloyd, Elisabeth. 2005. Units and Levels of Selection. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta.
URL: *http://plato.stanford.edu/archives/fall2005/entries/selection-units/*

Machamer, Peter, Lindley Darden and Carl F. Craver. 2000. "Thinking about Mechanisms." *Philosophy of Science* 67(1):1–25.

Machiavelli, Niccolo. 1958. *The Prince.* translated by W. K. Marriott. London: Everyman. First published 1515.

Mackie, Gerry. 1996. "Ending Footbinding and Infibulation: A Convention Account." *American Sociological Review* 61(6):999–1017.

McAdams, Richard H. 1997. "The Origin, Development, and Regulation of Norms." *Michigan Law Review* 96(2):338–433.

McClennen, Edward F. 1990. *Rationality and Dynamic Choice.* Cambridge et al.: Cambridge University Press.

McClennen, Edward F. and Scott Shapiro. 1998. Rule-guided behavior. In *The new Palgrave Dictionary of economics and the law*, ed. Peter Newman. London: Macmillan.

McNamara, Paul. 2006. Deontic Logic. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta.
URL: *http://plato.stanford.edu/archives/spr2006/entries/logic-deontic/*

Miller, John H. and Scott E. Page. 2007. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life.* Princeton University Press.

Mills, Susan K. and John H. Beatty. 1979. "The Propensity Interpretation of Fitness." *Philosophy of Science* 46(2):263–286.

Mäki, Uskali. 1992. "On the method of isolation in economics." *Poznan Studies in the Philosophy of the Sciences and the Humanities* 26:319–354.

Monroe, Kristen Renwick. 1997. Human Nature, Identity, and the Search for a General Theory of Politics. In *Contemporary Empirical Political Theory*, ed. Kristen Renwick Monroe. Berkeley and Los Angeles: University of California Press pp. 279–306.

Monroe, Kristen Renwick, James Hankin and Renee Bukovchik Van Vechten. 2000. "The psychological foundations of identity politics." *Annual Review of Political Science* 3:419–447.

Nowak, Martin A. 2006*a*. *Evolutionary Dynamics: Exploring the Equations of Life.* Cambridge (MA) and London: Belknap Press.

Nowak, Martin A. 2006*b*. "Five rules for the Evolution of Cooperation." *Science* 314:1560–1563.

Nowak, Martin A. and Karl Sigmund. 1998. "The Dynamics of Indirect Reciprocity." *Journal of Theoretical Biology* 194:561–574.

Nowak, Martin A. and Karl Sigmund. 2005. "Evolution of indirect reciprocity." *Nature* 437(7063):1291–1298.

Nowak, Martin A., Sebastian Bonhoeffer and Robert M. May. 1994. "Spatial Games and the Maintenance of Cooperation." *PNAS* 91(11):4877–4881.

Nowak, Martin and Karl Sigmund. 1993. "A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game." *Nature* 364(6432):56–58.

Ockenfels, Axel and Reinhard Selten. 2000. "An Experiment on the Hypothesis of Involuntary Truth-Signalling in Bargaining." *Games and Economic Behavior* 33(1):90–116.

Ohtsuki, Hisashi, Christoph Hauert, Erez Lieberman and Martin A. Nowak. 2006. "A simple rule for the evolution of cooperation on graphs and social networks." *Nature* 441:502–505.

Ohtsuki, Hisashi and Martin A. Nowak. 2006*a*. "Evolutionary games on cycles." *Proceedings of the Royal Society B* 273:2249–2256.

Ohtsuki, Hisashi and Martin A. Nowak. 2006*b*. "The replicator equation on graphs." *Journal of Theoretical Biology* 243:86–97.

Ohtsuki, Hisashi and Yoh Iwasa. 2006. "The leading eight: Social norms that can maintain cooperation by indirect reciprocity." *Journal of Theoretical Biology* 239(4):435–444.

Ones, Umut and Louis Putterman. 2007. "The ecology of collective action: A public goods and sanctions experiment with controlled group formation." *Journal of Economic Behavior & Organization* 62(4):495–521.

Osborne, Martin J. 2004. *An introduction to game theory.* New York: Oxford University Press.

Osborne, Martin J. and Ariel Rubinstein. 1994. *A Course in Game Theory.* Cambridge (MA) and London: MIT Press.

Pacheco, Jorge M., Arne Traulsen and Martin A. Nowak. 2006*a*. "Active linking in evolutionary games." *Journal of Theoretical Biology* 243(3):437–443.

Pacheco, Jorge M., Arne Traulsen and Martin A. Nowak. 2006*b*. "Co-evolution of strategy and structure in complex networks with dynamical linking." *Physical Review Letters* 97:025103.
URL: *http://arXiv.org/pdf/q-bio/0701008v1*

Page, Scott E. 2005. Agent Based Models. In *The New Palgrave Dictionary of Economics and Law.* 2nd ed. MacMillan.

Page, Talbot, Louis Putterman and Bulent Unel. 2005. "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency." *The Economic Journal* 115(506):1032–1053.

Pettit, Philip. 2000. "Rational choice, functional selection and empty black boxes." *Journal of Economic Methodology* 7(1):33–57.

Phillips, Michael. 1985. "Reflections on the Transition from Ideal to Non-Ideal Theory." *Nous* 19(4):551–570.

Railton, Peter. 2000. "Darwinian Building Blocks." *Journal of Consciousness Studies* 7(1-2):55–60.

Rawls, John. 1999*a*. *The Law of Peoples.* Cambridge, Mass. and London: Harvard University Press.

Rawls, John. 1999*b*. *A theory of justice.* Rev. ed. Oxford: Oxford University Press.

Rege, Mari and Kjetil Telle. 2004. "The impact of social approval and framing on cooperation in public good situations." *Journal of Public Economics* 88(7-8):1625–1644.

Reno, Raymond R., Robert B. Cialdini and Carl A. Kallgren. 1993. "The Transsituational Influence of Social Norms." *Journal of Personality and Social Psychology* 64(1):104–112.

Ridge, Michael. 2006. Moral Non-Naturalism. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta.
URL: *http://plato.stanford.edu/archives/sum2006/entries/moral-non-naturalism/*

Ruse, Michael. 1995. *Evolutionary Naturalism: Selected Essays.* London and New York: Routledge.

Sally, David. 1995. "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992." *Rationality and Society* 7(1):58–92.

Sally, David. 2000. "A general theory of sympathy, mind-reading, and social interaction, with an application to the Prisoners' Dilemma." *Social Science Information* 39(4):567–634.

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Santos, Francisco C. and J. M. Pacheco. 2006. "A new route to the evolution of cooperation." *European Society for Evolutionary Biology* 19:726–733.

Santos, Francisco C, Jorge M Pacheco and Tom Lenaerts. 2006. "Cooperation prevails when individuals adjust their social ties." *PLoS Computational Biology* 2(10):1284–1291.
URL: *http://dx.doi.org/10.1371/journal.pcbi.0020140*

Schelling, Thomas C. 1978. *Micromotives and macrobehaviour.* New York: Norton.

Segerstrale, Ullica Christina Olofsdotter. 2000. *Defenders of the truth: the battle for science in the sociobiology debate and beyond.* Oxford: Oxford University Press.

Selten, Reinhard. 2001. What is Bounded Rationality? In *Bounded Rationality: The Adaptive Toolbox*, ed. Gerd Gigerenzer and Reinhard Selten. Cambridge (MA) and London: MIT Press pp. 13–36.

Sen, Amartya. 2002. *Rationality and Freedom.* Cambridge (MA): Belknap Press.

Sen, Amartya K. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy and Public Affairs* 6(4):317–344.

Sethi, Rajiv and E. Somanathan. 2003. "Understanding reciprocity." *Journal of Economic Behavior & Organization* 50(1):1–27.

Sinnott-Armstrong, Walter. 1984. "'Ought' Conversationally Implies 'Can'." *The Philosophical Review* 93(2):249–261.

Skyrms, Brian. 1996. *Evolution of the Social Contract.* Cambridge; New York: Cambridge University Press.

Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure.* Cambridge: Cambridge University Press.

Skyrms, Brian and Robin Pemantle. 2000. "A dynamic model of social network formation." *PNAS* 97(16):9340–9346.

Sober, Elliott and David Sloan Wilson. 1998. *Unto others: the evolution and psychology of unselfish behavior.* Cambridge, Mass: Harvard University Press.

Spiekermann, Kai. 2007. "Translucency, Assortation, and Information Pooling: How Groups Solve Social Dilemmas." *Politics, Philosophy & Economics* 6(3):303–324.

Sterelny, Kim. 2006. "Memes Revisited." *British Journal for the Philosophy of Science* 57(1):145–165.

Streumer, Bart. 2003. "Does 'Ought' Conversationally Implicate 'Can'?" *European Journal of Philosophy* 11(2):219–228.

Strevens, Michael. 2007. "Why Explanations Lie: Idealizations in Explanation.". **URL:** *http://www.strevens.org/research/expln/Idealization.pdf*

Strogatz, Steven H. 2001. "Exploring complex networks." *Nature* 410:268–276.

Sugden, Robert. 2000. "Credible worlds: the status of theoretical models in economics." *Journal of Economic Methodology* 7(1):1–31.

Sugden, Robert. 2001*a*. "The evolutionary turn in game theory." *Journal of Economic Methodology* 8(1):113–130.

Sugden, Robert. 2001*b*. "Ken Binmore's Evolutionary Social Theory." *The Economic Journal* 111(469):213–243.

Teller, P. 2001. "Twilight Of The Perfect Model Model." *Erkenntnis* 55(23):393–415.

Tesfatsion, Leigh. 2007. "Agent-Based Computational Economics.".
URL: *http://www.econ.iastate.edu/tesfatsi/ace.htm*

Tomassini, M., L. Luthi and M. Giacobini. 2006. "Hawks and Doves on Small-World Networks.".
URL: *http://arXiv.org/pdf/physics/0612223*

Tooby, John and Leda Cosmides. 1992. The Psychological Foundations of Culture. In *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* New York: Oxford University Press pp. 19–136.

Trivers, Robert. 1971. "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46:35–57.

Vanderschraaf, Peter. 2006. "War or Peace? A Dynamical Analysis of Anarchy." *Economics and Philosophy* 22(2):243–279.

Voorhoeve, Alex. 2002. "The Good, the Right, and the Seemly: Ken Binmore Interviewed." *The Philosopher's Magazine* 21:48–51.

Weisberg, Michael. 2003. When Less is More: Tradeoffs and Idealization In Model Building PhD thesis Stanford University.

Wilson, David Sloan and Elliott Sober. 1994. "Re-introducing group selection to the human behavioral sciences." *Behavioral and Brain Sciences* 17(4):585–654.

Woodward, Jim. 2002. "What is a Mechanism? A Counterfactual Account." *Philosophy of Science* 69:S366–S377.