Does the Structural Evolution of Morality Lead to Moral Relativism?

Stefan Seil

Munich 2020

Does the Structural Evolution of Morality Lead to Moral Relativism?

Stefan Seil

Bachelor's Thesis
at the Faculty of Philosophy, Philosophy
of Science and the Study of Religion
at LMU Munich

submitted by
Stefan Seil

Munich, June 21st 2020

Contents

Chapter 1

Introduction

J. McKenzie Alexander's theory of the *Structural Evolution of Morality* (Alexander 2007) attempts to ground morality in a foundation of evolutionary game theory. According to the author, moral behaviour is a result of boundedly rational people attempting to maximize their expected utility in games of social interaction. These interpersonal decision problems take place in social networks which shape the outcome of the interactions. Alexander shows how common game-theoretic games like the Prisoner's Dilemma or the Stag Hunt could lead to the emergence of moral intuitions like cooperation and trust under the right pressure from evolutionary selection processes. Together with theories of bounded rationality and experimental psychology, the structural evolution of morality can give an explanation of our morality. One important question for the moral philosopher is thus what the implications of such an explanation of morality are. Superficially, it appears that the structural evolution of morality would render the truth of moral judgements entirely relative, thereby leading to moral relativism. Alexander briefly touches on this topic in his book, but ultimately doesn't make a conclusive argument about this topic. The goal of this paper is therefore to provide the missing investigation into whether the structural evolution of morality leads to moral relativism. The focus hereby lies on the descriptive part of moral relativism, in the sense of a diversity of moral beliefs along with substantial disagreements about those beliefs. The metaethical implications of this descriptive moral relativism are only going to be touched on at the end of the paper.

The paper is structured as follows. After the basic introduction in this chapter, chapter 2 serves to give an overview of the general topic. I present some background to evolutionary explanations of morality in section 2.1, and identify one important issue with such explanations that is relevant for the argument of the paper. Section 2.2 contains a brief introduction to the basic ideas of evolutionary game theory, as the structural evolution of morality is based on this theory. In section 2.3, I give a short overview of Alexander's theory including his main thesis and the methodology he uses for arguing for it. Section 2.4 is an attempt to provide a rough sketch of Alexander's conception of morality given the theory of the structural evolution of morality. In chapter 3, I introduce the issue of moral relativism with regards to the structural evolution of morality. Section 3.1 presents Alexander's argument about the implications of the theory on morality. I then differenti-

ate between descriptive and metaethical moral relativism in section 3.2, and introduce the concept of the domain of relativization, which is the anchor point for the argument of the paper. Section 3.3 describes the methodology of the descriptive analysis that is to follow. Chapter 4 represents the main argument of this paper, and it shows that the structural evolution of morality does indeed lead to descriptive moral relativism. I therefore derive the potential domains of relativization from Alexander's model in section 4.1. Following, I analyze the environmental and cognitive domains of relativization of the theory in sections 4.2 and 4.3, and investigate some additional convergence properties of the simulation in section 4.4. Section 4.5 summarizes the results of this chapter's argument. In chapter 5, I give a preliminary discussion of the metaethical implications resulting from the descriptive moral relativism, and from the nature of morality that the structural evolution of morality brings forward. Chapter 6 concludes the paper by summarizing the important points.

Chapter 2

The Structural Evolution of Morality

In this section, I am introducing the topic of this paper in broader terms. In section 2.1, I give some background to evolutionary explanations of morality and identify one important issue with such explanations, which is going to be important for the argument of this paper. Section 2.2 represents a brief introduction into the basic ideas of evolutionary game theory, as this is the theory that the structural evolution of morality is based on. In section 2.3, I give a short overview of Alexander's theory of the structural evolution of morality and introduce his main thesis as well as his methodology for arguing for it. Section 2.4 then describes a rough sketch of Alexander's conception of morality given its structural evolution, based on the limited information available.

2.1 Evolutionary Explanations of Morality

Considering that the human species has evolved through natural selection, we have to assume that evolutionary forces have shaped our morality at least to some degree. Investigations into the evolution of morality can be broadly categorized into two approaches (FitzPatrick 2016, sec. 1). One approach has mostly been of interest to philosophy, the other to science. For philosophy, the relevant problems arise in the realm of normative ethics and metaethics. These fields regard questions such as how evolutionary influences on our morality relate to the justification of certain moral norms, or to the truth status of moral judgements (see e.g. Joyce 2005). In science, the focus has been on providing explanations of morality, i.e. descriptive accounts of how evolutionary processes shaped or created our moral beliefs (see e.g. Shermer 2004). The results of this descriptive approach also have implications for the philosophical discussions. As the two approaches gain further insight into their respective fields of study, the gap between them is shrinking. The theory of the structural evolution of morality, which is the central topic of this paper, is one attempt to provide a descriptive explanation of our morality. It does not, however, take a traditional empirical approach as one might expect from the categorization above. Instead, it uses evolutionary game theory analyzed through agent-based computer simulations to provide a convincing account of how our moral intuitions came to be.

One important issue regarding the philosophical implications of evolutionary explanations of morality is that of *autonomous moral reflection*. It is sufficiently important for both the main investigation of this paper, as well as the metaethical implications touched on near the end of this paper, that this warrants explaining it here already. When talking about evolutionary explanations of morality, one needs to be aware of the distinction between reasons and causes. While evolutionary processes can give good explanations for our capacity to understand and discuss moral judgements, or even why we have certain basic moral intuitions guiding our lives, such explanations do not (necessarily) provide convincing explanations for our judgements of moral actions. When we are asking people why they acted in a certain way, we expect them to give us reasons. Together with their beliefs, we take those reasons to give an explanation for why they made the respective action. We do not usually take these actions to be fully causally determined by our genetics or some other factors influenced by our evolutionary history. When I choose to give money to a beggar on the streets, I am going to justify that action by saying that the beggar seemed to be in dire need of it, and that I believe that you ought to help people in need. I am not, however, going to justify the action by referring to the neurochemical processes in my brain, or the biological mechanism which lead my brain to develop to what it is at this point in time. Ethics, therefore, can be considered an independent system of theoretical inquiry, with its own methods and standards of justification, and it cannot reasonably be understood from the outside by appeal to explanations of the biological or physical phenomena which give rise to it (Nagel 2012, p. 142). We can summarise this notion through the following definition of autonomous moral reflection:

> "[People] have, to greater or lesser degrees, a capacity for reasoning that follows autonomous standards appropriate to the subjects in question, rather than in slavish service to evolutionarily given instincts merely filtered through cultural forms or applied in novel environments. Such reflection, reasoning, judgment and resulting behavior seem to be autonomous in the sense that they involve exercises of thought that are not themselves significantly shaped by specific evolutionarily given tendencies, but instead follow independent norms appropriate to the pursuits in question." (FitzPatrick 2016, sec. 2.4)

As we are going to see, the possibility of autonomous moral reflection is quite relevant for the discussions in this paper.

2.2 Evolutionary Game Theory

Evolutionary game theory provides the foundation of the theory of the structural evolution of morality. Giving a conclusive introduction into evolutionary game theory would be outside the scope of this paper. For a better coverage of the topic, see Jörgen W. Weibull's classic *Evolutionary Game Theory* (Weibull 1995). I do, however, want to briefly present the basic principles of the theory, because some of the points are going to be important for the discussion in this paper.

In short, evolutionary game theory can be viewed as the study of infinitely repeated interpersonal decision problems, faced by populations of boundedly rational individuals, whereby some evolutionary selection processes govern the behaviour of those individuals over time. We can take a classic example from evolutionary biology to make this clearer. Consider W. D. Hamilton's explanation for why populations of mammals have approximately equal rations of males and females (Hamilton 1967). The individuals of such a population face the interpersonal decision problem of reproduction, whereby each individual embodies the behaviour of having a tendency to produce more males, or that of having a tendency to producing more females. Here, the behaviour of an individual is fixed from birth, but this need not be the case in general. Let us now assume that the payoffs of the decision problem, i.e. reproduction, are the expected numbers of grandchildren an individual is going to have. If there are fewer males than females among the population, the behaviour of producing more males is then advantageous. A male has higher prospects for mating, which is going to lead him to create more offspring. His parents then get a higher payoff (expected number of grandchildren) in the decision problem of reproduction. The behaviour of producing more males is then going to spread genetically, until there is no relative advantage to producing males anymore. The same holds for females, respectively, when males dominate the population. Thus, over time, the evolutionary processes are going to move the population to a state where the amount of males and females in the population are approximately equal.

There are two main areas of interest to evolutionary game theory (Alexander 2019, sec. 2). For one, we can use the theory to analyze the stability of certain behaviours among a population at a specific point in time. The classical approach to this is the notion of an evolutionarily stable state, which specifies that a population of individuals with a certain distribution of behaviours can resist being taken over by individuals with novel behaviours. There can be other notions of stability as well, as we are going to see in section 4.4. Secondly, we can investigate the dynamics of the evolutionary processes, i.e. how the behaviours among the population change over time and whether they converge to some stable equilibrium. There are different models to do so. The classical approach lies in using the replicator dynamics, which analyzes how the frequency of certain behaviours changes among a large population in the limit. Another approach is using agent-based computer simulations, where we can model more sophisticated interactions and constraints compared to the replicator dynamics. As we are going to see in the next section, Alexander uses such an agent-based model for his theory of the structural evolution of morality.

While evolutionary game theory started as an application of game theory to evolutionary biology, it has become an especially valuable tool for economics and the social sciences. The analysis of the evolutionary processes has been carried over to explain market forces and cultural dynamics. Therefore, we can interpret the models used in evolutionary game theory in two ways. One is the interpretation of biological evolution, which was used in the example of the sex ratio among mammalian populations above. The other is the interpretation of cultural evolution, whereby we can view the behaviours as beliefs which are changed by the individuals over time. Here, it is clearer how the bounded rationality of the individuals comes into play, as the individuals then do in fact make choices on their

own. This interpretation is the main focus of Alexander's theory and the discussions in this paper. Note that there are certain problems which can arise from comparing the payoffs of different individuals in the cultural interpretation (see Grüne-Yanoff 2011). I am going to put these aside for now, as we are going to touch on this topic in section 4.2.

2.3 Alexander's Theory

J. McKenzie Alexander's work can be seen to follow a line of research in using evolutionary game theory to explain different social and moral phenomena. There has been a sizeable amount of work in this area already. In *The Evolution of Cooperation*, Robert Axelrod famously investigated cooperative behaviour in the repeated Prisoner's Dilemma by letting different strategies compete against each other in a computer-based tournament (Axelrod 1984). In his two-volume work *Game Theory and the Social Contract*, Ken Binmore advocates for evolutionary game theory as a systematic tool for investigating ethics, and uses the theory to argue about moral and political philosophy in the tradition of Hume, Rawls and Harsanyi (Binmore 1994; Binmore 1998). Brian Skyrms analyzes moral and social phenomena such as justice and altruism with evolutionary game theory in his work *Evolution of the Social Contract*, where he uses the replicator dynamics to investigate the evolutionary dynamics of these behaviours (Skyrms 2014). Alexander's addition to this lineage is that he uses agent-based models to investigate the structural properties of a variety of interpersonal decision problems, which can lead to the emergence of some common moral intuitions we have today.

In *The Structural Evolution of Morality*, Alexander attempts to ground morality in a foundation of evolutionary game theory. The main thesis of the book can be summarized as follows: Moral behaviour is a result of boundedly rational people attempting to maximize their expected utility in games of social interaction, whereby these games take place on social networks which themselves shape the outcomes of the interactions. In order to argue for this point, he analyzes four different game-theoretic games which contain strategy profiles that are supposed to model different moral intuitions. The Prisoner's Dilemma can be seen to reflect *cooperation* when players choose the Pareto-efficient outcome of both playing Cooperate. The Stag Hunt models a game of *trust*, because players need to trust each other to choose Stag in order to get the highest payoff. The Nash bargaining game (also called divide-the-cake) prompts *fairness*, as players can choose an equal split which leaves everybody with the same payoff and doesn't waste any of the resource. The ultimatum game offers the possibility of *retribution*, when a player rejects a seemingly unfair demand even though it is sub-optimal in the short-term. Alexander thus attempts to show through evolutionary game theory that the strategies reflecting these moral intuitions can be selected by evolutionary pressures as the dominant behaviour among a population. This can then offer an explanation for why we hold these moral norms of cooperation, trust, fairness or retribution. For the most part, his attempt succeeds. The only exception is the ultimatum game, where he could not convincingly show how strategies encoding a norm of retribution could reliably succeed in the model (see Alexander 2007, ch. 6).

As mentioned earlier, Alexander uses agent-based simulations to analyze the evolutionary dynamics of the social interactions. These simulations provide a more sophisticated model compared to the replicator dynamics, in that they offer more fine-grained control over the interactions between the individuals, as well as the possibility to add certain constraints (see ibid., sec. 2.2). A typical simulation of Alexander's model can be seen to work as follows: A population of agents is randomly initialized with strategies. The agents populate a social network which determines the kind of interactions that can take place. When interacting with another agent, an individual gets a certain payoff respective to the game that is being played. After a round of interactions, an agent can update his strategy by imitating a neighbor who got a better payoff than himself. This way, the distribution of strategies changes over time. The most important difference to simpler models of evolutionary dynamics is the addition of the structural component, hence the *structural* evolution of morality. Whereas the replicator dynamics model assumes that every agent has an equal probability of interacting with any other agent, the social networks used in Alexander's agent-based model constrain the possible interactions by virtue of their network topology. There are four different kinds of network topologies which are investigated in the book (see ibid., sec. 2.2–2.5). Lattice models constrain agents on a grid, whereby an individual can only interact with the individuals surrounding his cell. Small-world networks model a social network in which the majority of people have few social relations, while a handful of well-connected individuals exhibit additional relations across the network. These additional "bridge edges" effectively shorten the path between many of the other agents. Bounded-degree networks limit the amount of relations each agent can have through a lower and an upper limit. Dynamic networks offer the possibility of changing one's social relations over time, which leads to the social network being influenced by the interactions taking place inside of it. The crucial idea is that these different network topologies have different influences on the evolutionary dynamics. While the Prisoner's Dilemma analyzed through the classic replicator dynamics necessarily leads to a state of complete defection (ibid., pp. 56–59), a one-dimensional lattice-based model can result in Cooperate being selected as the dominant strategy (ibid., pp. 71–73). The evolution of morality is therefore *structural*, in that it is a consequence of the structural constraints on the interactions between individuals.

Let us make one concrete example in order to make it clear how the structural evolution of morality is supposed to work. Consider a population of people facing interpersonal decision problems corresponding to the Prisoner's Dilemma. The social relationships of the people are structured in such a way that most of the persons have few acquaintances, while a few individuals have relationships with more people. The social network which describes these relationships can thus be thought of as a small-world network. Each person among the population has a strategy for the Prisoner's Dilemma (Cooperate or Defect), which he uses for his interactions. The people now regularly engage in Prisoner's Dilemmas with their respective acquaintances. When a person has finished an interaction, he gets a payoff corresponding to the payoff structure of the game. The person then investigates his social relationships and looks for someone who got a better payoff than him in the previous interactions. If he finds such a person, and this person was using a different strategy than

him, then he chooses to imitate this strategy for the upcoming interactions. This way, the strategy of an individual changes. Over time, this strategy updating mechanism will lead to the population converging to a stable state, whereby the changes in strategies have come down to a minimum. The evolutionary dynamics have thus selected an equilibrium, which can contain one dominant strategy that the majority of people are using. Depending on a number of factors, the dominant strategy of the population can turn out to be Cooperate or Defect. Together with the mechanisms alluded to in the following section, this stable strategy can then, over time, turn into a moral intuition stating that cooperation is good (when Cooperate is dominant) or bad (when Defect is dominant).

2.4 Alexander's Conception of Morality

According to Alexander's theory, morality can be ultimately defined as follows: Moral principles are heuristics for maximizing expected utility in the long run, which are shaped by evolutionary forces governing interpersonal decision problems. Unfortunately, this account leaves many further questions unanswered. We are going to come back to these missing parts of the theory in chapter 5. Hence, I am going to reconstruct Alexander's conception of morality based on the limited information available.

First, we need to differentiate actions from behaviour. An action requires an intentional state, while moral behaviour doesn't. Alexander then makes a distinction between *thin* and *thick* descriptions of morality (Alexander 2007, p. 268). A person is said to thickly conform to a moral norm if she makes an action (intentional) and holds the necessary emotional states that motivate her action. She then holds "sufficiently many of the beliefs, intentions, preferences, and desires" (ibid.) that we typically require for the action to be considered a moral action. Conversely, thin conformity to a moral norm is an action which lacks the aforementioned emotional components. On its own, the model described in the book only delivers a thin description of morality. The condition of the intentionality on part of the individuals can be considered fulfilled under an interpretation of cultural evolution (ibid., p. 269), but the condition of emotionality is missing. In order to provide a thick description of morality, Alexander suggests we need to turn to a combination of different theories: evolutionary game theory, bounded rationality, and experimental psychology (ibid., pp. 274–277). The theories each play a different role in the thick explanation of morality. Evolutionary game theory describes the structure of the social interactions we face in our lives. Bounded rationality delivers tools for coming up with effective strategies for this structure (as the more realistic alternative to finding perfectly rational solutions). Experimental psychology (e.g. in the form of evolutionary psychology) can describe the missing emotionality which motivates individuals to act according to the respective strategies. Moral principles can then be considered fast and frugal heuristics for boundedly rational individuals. These heuristics also serve to limit the vast search space described by the evolutionary game theory, in order to face complex interpersonal decision problems without having to consider all possible options.

The combination of these three theories can in principle deliver a thick description

of morality. But what about more sophisticated moral norms than those covered in the book? To answer this question, Alexander makes a distinction between *natural* and *artificial* virtues (which he adopted from Hume, but reinterpreted to have slightly different meanings) (ibid., pp. 282–286). Natural virtues are those norms which are the product of the evolutionary pressures, they are hard-wired and not chosen by us. Artificial virtues describe the norms which we have chosen and created ourselves. We can interpret the moral principles which are a result of the evolutionary dynamics described in the book as being natural virtues. The crucial remaining question is in how far the natural and artificial virtues influence each other. Alexander ultimately leaves it open in how far the natural virtues constrain the artificial ones. One paragraph gives some insight into the other direction:

> "I suspect that normative discussion plays a less significant role in our judgments of fairness. I think our sense of fairness is fairly well calibrated to track behaviors and outcomes that satisfy our preferences to the greatest extent possible subject to constraints. The theories and principles which we use to explain why an outcome is fair might very well be shaped by normative discussion (theories have to come from somewhere), but the reason why those theories have the form that they do is not due to discussion. Moral theories have the form they do because it is a fact that social beings such as us, who have preferences of a certain kind, maximize our long-run expected utility by behaving in ways that conform to certain moral principles." (ibid., p. 278)

While he specifically talks about fairness here, we can extend this statement to cover the other moral principles covered in the book as well. I interpret this paragraph to state the following: The natural virtues are determined by the evolutionary dynamics we faced in the past. Any rational deliberation as part of a normative discussion about these norms is a mere *post-hoc rationalization*. The normative discussions (and thus the artificial virtues which could originate from them) do not have the capacity to change the natural virtues in any way, they can only attempt to give justifications of them after they are already in place. If this interpretation is correct, it has a big impact on an individual's capability of autonomous moral reflection. Alexander's explanation of morality then essentially denies that this is possible, at least for those moral principles (i.e. natural virtues) covered in the book. For example, a normative discussion about fairness might lead us to come up with reasons for why fairness is good, which we can then use to justify a moral action. The discussion cannot lead us into believing that fairness is bad, though, because this norm is ingrained in us as an evolutionary adaptation.

Chapter 3

Moral Relativism

In this chapter, I introduce the issue of moral relativism with regards to the structural evolution of morality. In section 3.1, I present Alexander's argument about the implications of his theory on morality. Section 3.2 serves to make the important distinction between descriptive and metaethical moral relativism, and to introduce the concept of the domain of relativization, which is going to be the basis for my investigation into descriptive moral relativism. In section 3.3, I finally describe the methodology of the analysis that is going to follow in chapter 4.

3.1 Alexander's Argument

At the end of his book, Alexander writes one paragraph which touches on the implications of his theory with regards to moral relativism:

> "[1] This means that our moral beliefs are simultaneously relative to our evolutionary history and our cultural background, but at the same time objectively true. [2] Insofar as our moral beliefs provide solutions to interdependent decision problems, we cannot say that any one solution is better than any other — in an abstract sense — because, detached from our preferences, there is no absolute standard from which to judge. [3] Given our preferences, and from our own personal point of view, there can be an objective moral theory that prescribes the best way of satisfying those preferences." (Alexander 2007, p. 291)

This paragraph includes three different statements, which I have marked accordingly in the quote above. The first statement hints at an important phenomenon of the structural evolution of morality. Depending on how the evolutionary processes play out, different strategies are selected for, which again lead to different moral norms being created. This can be seen as a merely descriptive matter of fact. The second statement deals with the judgement of moral actions. Consider the situation in which the evolutionary forces lead to two different dominant strategies in two different societies. Inside the first, the stable and utility-maximizing strategy is playing Defect every time, while inside the second it is

playing Cooperate every time. For the individuals in the first society, the strategy that they employ is better than the one from the second society, and vice versa. Every individual follows the best strategy available to them. Thus, from the standpoint of the observer, it cannot be said that playing Defect is better or worse than playing Cooperate *in general*. This hints at an argument against moral objectivism. One notion that could be argued for is that moral judgements don't have any truth value altogether (e.g. in the form of moral non-cognitivism). Another, more obvious position is that moral judgements are simply relative to certain properties. In the case of the above example, they are relative to the society an individual lives in. Alexander's statement does not in itself make a decision as to which position should be chosen. However, a video from the London School of Economics about Alexander's work hints at moral relativism being the author's preferred notion (see Alexander 2015).

The third statement may seem contradictory at first. Again, the video just mentioned can help to clear up any misunderstandings. Alexander uses two different dichotomies: relative vs. absolute and subjective vs. objective. Potential confusion arises from the fact that, in metaethics, objective and relative are often taken as opposites with regard to the truth status of moral judgements. Subjectivism is often considered as a special case of relativism, and the term absolute is not found as frequently in the metaethical literature. Using Alexander's (arguably more correctly employed) terminology, moral norms are considered to be relative (instead of absolute) in that they are context-specific as described above. Additionally, they are objective (instead of subjective), because "there is a fact of the matter over what behaviour maximally satisfies your preferences given the constrains placed by other people" (ibid.). The latter notion of objectivity is more commonly referred to as *moral realism*.

This short paragraph obviously doesn't do the matter justice. Therefore, the goal of this paper is to provide the missing argument about the implications of the structural evolution of morality on moral relativism. Considering that moral relativism is my focus, I am going to put the third statement aside and assume that moral realism holds.

3.2 Descriptive and Metaethical Moral Relativism

When talking about moral relativism, there are two different forms of arguments which need to be differentiated: *descriptive* and *metaethical* moral relativism. Descriptive moral relativism is a purely descriptive statement about matters of fact. We can take the following definition as an outline:

> "As a matter of empirical fact, there are deep and widespread moral disagreements across different societies, and these disagreements are much more significant than whatever agreements there may be." (Gowans 2019, sec. 2)

The moral disagreements which are at the heart of this definition have to be understood empirically: Two individuals disagreeing with each other, whereby the claims made are in some way incompatible with each other. The qualitative notion of the disagreements

being more significant than the agreements makes this definition somewhat fuzzy. When arguing for descriptive moral relativism, it helps to make it plausible why the disagreements dominate over the agreements.

Metaethical moral relativism is a normative statement about the truth status of moral judgements. We can again take some orientation from the following definition:

> "The truth or falsity of moral judgments, or their justification, is not absolute or universal, but is relative to the traditions, convictions, or practices of a group of persons." (ibid., sec. 2)

The exemplary listing of traditions, convictions and practices is of course not conclusive. In principle, we can make an argument that the truth of moral judgements is relative to any conceivable thing. In this regard, we can make use of some terminology from the philosophical literature on relativism (which includes, but is not limited to *moral* relativism). Let us call that which the truth status of a moral judgement or its justification is relative to a "domain of relativization" (Baghramian and Carter 2019, sec. 1.1). In the definition above, traditions, convictions and practices would each be a potential domain of relativization.

There is one important point to notice at this point. While the relation of something being "relative to" something else only occurs in the definition of metaethical moral relativism, the domains of relativization are already present in the descriptive notion of moral relativism. They appear as correlations between variables in the data set. Let us consider an example to make this clear. Consider an anthropologist who is charged to analyze whether descriptive moral relativism holds. This person will collect data around the world and assess whether people more often than not disagree about moral judgements. If this person had an exhaustive data set about all the characteristics of the people under investigation, including their history, their culture, and so forth, then one could attempt to uncover possible domains of relativization through data analysis. Given an interaction between two people, which variables in the data set does a disagreement between these two people correlate with? It could, for example, correlate with differences in cultural background of the two individuals. One possible domain of relativization would thus be the cultural background of a person making a moral judgement. Note that this does not make any normative assumption about the truth status of the differing moral judgements. It merely offers some possible domains of relativization which, when plausible, can be used to argue for metaethical moral relativism.

Most discussions of metaethical moral relativism begin with the assumption that descriptive moral relativism is true (Gowans 2019, sec. 4). Thus, for the purposes of this paper, I assume that descriptive moral relativism is a necessary condition for metaethical moral relativism. Is it, however, not a sufficient condition. Given descriptive moral relativism, there is a variety of non-objectivist metaethical positions one could argue for. Such alternative notions include moral skepticism (the idea that we cannot know the truth status of moral judgements) and moral non-cognitivism (the idea that moral judgements do not have relevant truth states at all) (see ibid., sec. 6). The task of proponents of

metaethical moral relativism is thus not only to argue against moral objectivism, but to make a convincing point as to why this is the best way to view the truth status of moral judgements. That is, there needs to be a convincing argument for the following thesis: As a matter of fact, moral judgements have a knowable truth status, and the best way to think about this truth status is that it is not absolute, but relative to some domain of relativization. Making a convincing argument for metaethical moral relativism then requires being specific about what the domains of relativization for the truth status of moral judgements are. Notice that descriptive moral relativism, being a necessary condition, needs to be established beforehand. One can therefore make use of the domains of relativization offered by the descriptive analysis for the metaethical argument.

3.3 Methodology

Due to the limited scope of this paper, and because of the missing information about Alexander's conception of morality, I can not provide a conclusive argument for or against metaethical moral relativism in this paper. I therefore focus on showing that the structural evolution of morality leads to descriptive moral relativism. An argument for the metaethical position can then use the domains of relativization provided by the descriptive analysis in this paper to make a strong point. Following, I want to specify some basic assumption as well as the methodology for the upcoming analysis of descriptive moral relativism.

Remember that descriptive moral relativism merely describes the matter of facts. If we take the structural evolution of morality to provide an explanation of *our* morality (i.e. the moral beliefs that humans hold right now), then we could simply refer to existing analyses of the moral diversity in our world. If such a literature review lead to descriptive moral relativism being regarded as true, then the question of whether the structural evolution of morality leads to moral relativism would have to be answered positively. Otherwise, we would need to assume that the structural evolution of morality is false, in that it does not provide an explanation for our morality, because it could not explain the variety of different moral beliefs characterized by descriptive moral relativism. This obviously misses the point of the exercise. We therefore need to argue, in a sense, counterfactually. We need to investigate not the current state of our world, but rather whether *a world* governed by the evolutionary forces described by the structural evolution of morality would lead to descriptive moral relativism.

In doing so, we need to underline the interpretation of the possibility of autonomous moral reflection from section 2.4. I interpreted Alexander's statements to say that the moral norms covered in the book, which are products of the evolutionary forces governing interpersonal decision problems, cannot be revised through normative discussion after they have been selected for. In order to investigate descriptive moral relativism from the standpoint of the structural evolution of morality, we need to make the assumption that this interpretation is correct. Remember that descriptive moral relativism concerns disagreements about moral beliefs. If autonomous moral reflection about the adaptive moral norms was feasible, then it would be possible for individuals to overwrite these norms through

normative discussion. The moral beliefs held by individuals (and thus the disagreements which come up between them) would not be causally determined by the structural evolution of morality anymore, and we couldn't argue about them from the standpoint of the theory. Let me make an example to make this point clear: Consider again the situation in which the structural evolution of morality leads to two stable strategies in two different societies. One is governed by people playing Defect in the Prisoner's Dilemma, the other by playing Cooperate. Individuals living in the former society might hold the moral belief that cooperation is wrong. Imagine that those individuals now, through autonomous moral reflection, overwrite their adaptive moral belief to state that cooperation is right. When we now analyze the diversity of moral beliefs in this hypothetical world (encompassing both societies), we will not find any disagreements anymore. Therefore, we must assume that it is not possible to change these adaptive moral beliefs through autonomous moral reflection.

How do we analyze whether the structural evolution of morality leads to descriptive moral relativism? As mentioned in the previous section, it helps to derive possible domains of relativization during the descriptive investigation, which could also later be used for a metaethical discussion. This can be done in a top-down or bottom-up manner. For the top-down approach, we can think of plausible domains of relativization which we then analyze in the context of the structural evolution of morality. Philosophical literature on relativism can offer suggestions in this regard. Possible domains could include language, culture, community or individuals (Haack 1996, p. 297). Unfortunately, many of these domains provide to be too high-level for our analysis. Therefore, it is better to follow the bottom-up approach. We thereby look at the model Alexander uses to argue for the structural evolution of morality, and extract domains that are meaningful in the context of that model. The analysis of descriptive moral relativism can then be carried out along those domains. I am going to derive the possible domains of relativization in the structural evolution of morality at the beginning of the next chapter.

Chapter 4

Descriptive Analysis

In this chapter, I show that the structural evolution of morality leads to descriptive moral relativism. In section 4.1, I derive the potential domains of relativization from Alexander's model. Sections 4.2 and 4.3 then deal with the analysis of the domains belonging to the environment of the social interactions and to the cognition of the individuals, respectively. In section 4.4, I analyze some additional domains of relativization regarding the convergence of Alexander's model. The results of the analysis are summarized in section 4.5.

4.1 Domains of Relativization

We want to investigate whether the structural evolution of morality leads to descriptive moral relativism. As mentioned in section 3.3, we can approach this by analyzing the potential domains of relativization within Alexander's theory. Specifically, we take a bottom-up approach and derive these domains from the agent-based model itself. As a point of reference, we can take a look at the different stages which a simulation of Alexander's model goes through:

1. **Initialization:** At the beginning, the simulation is initialized with a default distribution of strategies among the population of agents.

2. **Network:** The agents engage in interpersonal decision problems. These interactions are determined by the topology of the social network.

3. **Payoff Structure:** During each interaction, a predefined game-theoretic game with a certain payoff structure is played.

4. **Learning:** Depending on how an agent's payoff compares to the payoffs of his neighbors, the agent can change his strategy for subsequent interactions according to a learning rule.

5. **Randomness:** Throughout the simulation, agents can be allowed to randomly experiment with different strategies, adding a probabilistic component to the model.

6. **Evaluation:** At some point in time, the simulation is stopped and the distribution of strategies among the population is evaluated.

These six components can serve as a good first approximation of possible domains of relativization. We can assume that each of the components' inner workings is determined by one or multiple parameters. If different values for one parameter lead to different dominant strategies at the evaluation of the simulation, then whatever this parameter reflects is a good candidate for an actual domain of relativization. For example, if the kind of learning rule used in the model influences the strategy which is going to be selected, then the part of an individual's cognition which is modelled by this parameter could be considered a viable domain of relativization. For purposes of improved clarity, I group the aforementioned components of the simulation into three categories, each corresponding to one of the following three sections. *Environment* deals with the initial distribution of strategies, the network topology and the payoff structure of the games. *Cognition* concerns the strategy updates of the agents in the simulation. *Convergence* regards randomness as well as the time of evaluation.

As we have seen in section 3.2, in order for descriptive moral relativism to be true, the disagreements about moral beliefs resulting from the structural evolution of morality need to be *more significant* than the possible agreements. The sophisticated way to argue for this would be through a statistical analysis correlating the results of the simulations with their parameters. This would require access to the model, though, which is why I have to make the argument by referring to Alexander's experiments described in his book. Given this circumstance, I cannot exhaustively list all possible correlations. Doing so would (if possible at all) also amount to nothing more but a reformulation of large parts of Alexander's book. In order to make the argument as strong as possible with these limitations in mind, I define the question I want to answer in this chapter as follows: Do *plausible* changes to the parameters of the model *reliably* lead the model to produce *substantially* different results? The changes need to be *plausible*, so that we can be reasonably sure that the phenomena modeled by the simulations could realistically have occurred in the real world. While I cannot provide conclusive empirical evidence for this, I am at least going to argue that they are not completely unrealistic. The changes need to *reliably* lead to different results, so that the disagreements in moral beliefs created by the structural evolution of morality are not just rare coincidences where the conditions happened to be unfavorable for creating common moral intuitions. I am going to illustrate this by showing that many different parameters can change the outcome of the models, thereby offering a large potential for ensuing moral disagreements. The differences in results need to be *substantial*, in that Alexander's theory leads to a variety of different conflicting moral beliefs. Considering that the decision problems covered in the book do not have many strategies available to them in the first place, the way I am going to argue for this is to show that different results can emerge for *all* of these games. The selection of components and parameters mentioned above offers a good basis for the substantiality of the moral disagreements created by the structural evolution of morality. In the next three sections, I am going to show how these different parameters reliably lead to different results, and discuss whether differences in

their values are plausible.

4.2 Environment

The model used for analyzing the structural evolution of morality contains a number of parameters which could be labeled as belonging to the *environment* of an individual. These parameters include the initial distribution of strategies, the social network topology as well as the payoff structure of the games being played.

At the beginning of the simulation, the population of agents needs to be initialized with a distribution of strategies. This is usually done randomly. The initial distribution can have a decisive effect regarding the possibility of certain strategies becoming dominant. For the replicator dynamics, it is obvious that the results are completely determined by the initial state of the population. But agent-based models with more sophisticated network topologies can show this behaviour, too, even when allowing for experimentation. There are a number of examples to be found. In the Prisoner's Dilemma played on a one-dimensional lattice, the survival of cooperative clusters is fully determined by the initial distribution of strategies (Alexander 2007, p. 68). It also has a large effect on the outcome of the model in small-world networks (ibid., p. 83). For the Stag Hunt played on a one-dimensional lattice, the initial amount of stag and hare hunters can also completely determine the outcome of the simulation (ibid., pp. 114–117). In dynamic social networks, the evolutionary processes reliably select Stag as the dominant strategy as long as the initial distribution contains at least two stag hunters (ibid., pp. 145–146). Similarly, the Nash bargaining game on a two-dimensional lattice can only lead to fair division if enough people initially play the respective fair strategies (ibid., p. 175). More generally, we can see that any strategy which is supposed to be selected as dominant needs to be present initially in the population, unless we allow for experimentation. Probabilistic effects mutating the strategies of the individuals can thus relax the need for a specific composition of strategies at the start of the evolutionary process.

The network topology can also play a big role in determining the outcome of the simulation. For one, we have to differentiate between static and dynamic networks. The kind of dynamic networks used by Alexander are adopted from a model for social network formation by Brian Skyrms and Robin Pemantle (Skyrms and Pemantle 2000), in which agents can adjust the interaction probabilities for different connections based on the payoffs received from past interactions. This has the effect that connections between certain agents can be removed (almost) completely as the simulation progresses. In the Prisoner's Dilemma, this can lead to Cooperate being selected as the dominant strategy every time, as long as a few other requirements hold (Alexander 2007, pp. 98–100). But even among the static network topologies, there are significant differences to be found. For the Prisoner's Dilemma, it is impossible that Cooperate emerges as the dominant strategy in certain one-dimensional lattices (ibid., p. 71). In bounded-degree networks, it happens only rarely, depending on the chosen degrees (minimum and maximum number of edges per node) (ibid., p. 91). On a two-dimensional lattice, cooperation can emerge, again depending on

other parameters (Alexander 2007, p. 73). The result of the Nash bargaining game played on a bounded-degree network is also strongly influenced by the number of edges per node (ibid., pp. 194–195). In the Ultimatum game, the way the interactions are structured on a two-dimensional lattices greatly benefits unfair strategies (ibid., pp. 227–228). Small-world networks, however, can offer some protection from unfair strategies taking over the population (ibid., p. 231).

Different games are obviously going to lead to different outcomes. This is not problematic, as long as we accept that the moral intuitions which are the product of the structural evolution of morality are bound to reflect the solution to one specific interpersonal decision problem. The only possible caveat to this are the Prisoner's Dilemma and the Stag Hunt. The infinitely iterated forms of these two games can, under certain conditions, turn into each other (ibid., p. 110). In such a case, it would theoretically be plausible to apply a norm of cooperation to the Stag Hunt, or a norm of trust to the Prisoner's Dilemma. If the evolutionary processes selected Stag and Defect, or Cooperate and Hare as the dominant strategies, this could not only lead to additional disagreements about moral beliefs, but also to inconsistencies among the moral beliefs of one singular individual. However, considering that this is a very special case, we are going to grant that it does not happen reliably enough to pose a problem. What does reliably change the outcome of the evolutionary processes, though, is the payoff structure of a game. In the Prisoner's Dilemma, the emergence of cooperation is dependent upon the exact payoff structure in all of the available social network topologies (ibid., pp. 71, 73, 83, 91, 98–99). This effect is also apparent in the Stag Hunt, where the choice of the payoff matrix relates to which strategy is risk dominant (ibid., p. 103). On a one-dimensional lattice, this determines the outcome of the simulation (ibid., pp. 116–117). In small-world networks, it determines whether regions of hare hunters can survive the evolutionary dynamics (ibid., p. 131). For the Nash bargaining game and the Ultimatum game, there is less room for interpretation. Since the payoffs describe a portion of the available resources in these games, we can expect the actual numbers used to not be relevant, as long as the strategies are evenly divided with regards to the size of the resource.

Before we discuss the plausibility of the parameters, we have to make an observation about the differences in possible payoff structures. As we have just seen, different value assignments to the abstract payoffs of a game can influence the result of the simulations. The specific payoffs of an interpersonal decision problem thus have an impact on the extent of the moral disagreements produced by the structural evolution of morality. How can we interpret the different choices in payoff matrices? Considering that different payoff structures can lead to different strategies being selected, is it still plausible to think of moral beliefs mapping to abstract game-theoretic games? Consider the following example: Two interpretations of the Stag Hunt are each played among two different populations. In one population, the evolutionary dynamics lead to Stag and Hare being selected for the first and the second interpretation, respectively. In the other population, the dominant strategies are selected the other way around, with Hare for the first and Stag for the second interpretation. Those outcomes could lead to moral disagreements about different interpretations of the same game among the two populations. The proper level of abstraction is thus possibly

not the abstract game-theoretic game, but rather the specific interpretation of its payoffs. The implication for descriptive moral relativism is apparent: Since it is now possible to disagree about the right way to act in a specific interpretation of a game, there is even more room for disagreements.

The above examples show that changes in the three environmental parameters can reliably lead to different results. We can now turn to the question of plausibility. Are differences in the values of these three parameters plausible, in that they could reasonably occur in the real world? In the simulations, the initial strategies of the individuals are distributed randomly. In the real world, we cannot expect that a population of individuals spontaneously comes together and randomly distributes strategies among the individuals. However, there is not much more that we can figure out from the standpoint of the structural evolution of morality. The initial strategy is already set before the strategy updating mechanism kicks in. Therefore, it is determined by some factor outside of the model. Without any further insight into this part of the evolutionary process, and without making further assumptions, it is indeed best to view the initial distribution as random. From this viewpoint, the selection of a particular strategy would therefore in part be determined by luck.

With regards to the network topology, I argue that the differences in social network structure seem plausible. The contrast between a hunter-gatherer society and a modern city should make this apparent. Of course, one will be hard-pressed to find an example of a society which is structured according to a uniformly inhabited grid, whereby each individual only ever interacts with his direct neighbors. However, the important point here is that real social networks need not necessarily look like the ones used in the simulations. The lattice models should merely be seen as the simplest way to model spatial constraints for the interactions of the individuals. The point is that *differences* in social network topologies can have an influence on the emergence of certain dominant strategies, not that the networks need to have a specific structure. One might raise the objection that real-world social networks are always dynamic in nature, and that we can therefore ignore the static topologies. As mentioned earlier, the dynamic networks used in the model tend to rather quickly turns into a state where the interaction probabilities have converged towards one and zero. This way, for all practical purposes, a dynamic network turns into a static one. We can then view it as just another static network topology.

For the payoff structure of the games being played, the plausibility depends on what we interpret the payoffs structure to mean. In a biological interpretation of the evolutionary dynamics, the individual payoffs can be seen to correspond to expected Darwinian fitness, or changes therein. Considering that we focus on a cultural evolutionary interpretation, though, this interpretation of a payoff structure does not help us much. It seems far-fetched to say the we base our decision on how well they maximize our expected fitness. What do the payoffs correspond to, then? Alexander views cultural evolution as simply describing the change of beliefs over time (ibid., p. 19), and he keeps the interpretation of payoffs somewhat abstract by viewing them as reflections of the expected "satisfaction of our personal preferences" (ibid., p. 22). This can be seen as conform with traditional game theory. Note, however, that the actual numbers of the payoffs don't have any special

meaning in traditional game theory. They only reflect the order of a person's preferences. Preferences themselves are considered to be not comparable among persons. In evolutionary game theory, however, we implicitly assume that comparisons like these are possible. The learning rules used by Alexander's model (discussed in the next section) compare the payoff numbers of one agent with that of another agent. By using such a process, we have effectively already made the assumption that interpersonal comparisons of preferences are possible. Thus, to sum up the point, the payoffs can be viewed as comparable measures of the degree of satisfaction for an individual's preferences. In this light, it seems plausible that different instances of a game-theoretic game provide different amounts of preferential satisfaction to the players. As long as we accept this interpretation, differences in payoff structures therefore pose no problem to the plausibility criterion.

4.3 Cognition

There are some parameters of the model which I subsume under the category of *cognition*. They deal with the cognition of the individuals insofar as they lie within those individuals themselves, and are not determined only by environmental factors. Considering that we focus on the cultural interpretation of evolutionary dynamics, and that therefore the agents of the simulations represent boundedly rational individuals, there is no problem in justifying some basic cognitive abilities which can give rise to these parameters. The parameters I cover in this section are the different learning rules for updating an agent's strategy, a radius of interaction for the game-theoretic interactions and the strategy updates, as well as a variety of additional parameters which appeared infrequently in Alexander's experiments.

Alexander defines four different learning rules, which specify how an agent updates his strategy after a series of interactions: Imitating the best neighbor (*Imitate the Best*), imitating neighbors with probability proportional to their success, imitating the neighbor with the best average payoff, and choosing a strategy which would lead to the best response if neighbors keep using their current strategies (*Best Response*) (Alexander 2007, pp. 39–41). For most of the simulations described in the book, Imitate the Best is used as the learning rule governing the strategy updates. There are a handful of cases, though, where different rules were investigated and lead to different outcomes. For the most part, this happens in the Stag Hunt. As mentioned in the previous section, which strategy in the Stag Hunt is risk dominant can have a substantial influence on which strategy is going to be selected. There is a connection between the specific payoff structure of the game and the convergence of the model under influence of random mutation when using the Best Response learning rule (ibid., pp. 125–126). If Hare is risk dominant, then the interactions on a one-dimensional lattice can lead the population to converging on Stag under imitation, but fail to do so under best-response learning (ibid., pp. 121–122). Similarly, on the two-dimensional lattice, Best Response makes it very hard for Stag to spread among the population (ibid., pp. 127–128). Both on small-scale networks and bounded-degree networks, best-response learning prevents the spread of Stag unless the payoffs are chosen in a specific way to favor stag hunters (ibid., pp. 131–133, 138). Influence of the learning

rule can also be witnessed in the Nash bargaining game played on a two-dimensional lattice. When using Imitate the Best, a population of individuals mostly using an equal split strategy is very stable under mutations. For Best Response, though, the simulation will generally not reach a state of dominating equal-split strategies to begin with, and even if it does, mutations lead those fair strategies to being driven out over time (ibid., pp. 180–182). The learning rules also contain a *tie-breaking rule*, which governs what strategy to adopt in cases of equally good choices. The way this rule is defined can also have an influence on the outcome of a simulation, even though we can only find one example of this. When playing the Prisoner's Dilemma on a one-dimensional lattice, the rule determines whether cooperative regions can resist being taken over by defectors (ibid., p. 65).

The interaction and learning radius of an individual also affects the behaviour of the model on static network topologies. The learning radius governs how many other agents an individual takes into account when updating his strategy through imitation or best-response learning. The interaction radius defines the neighborhood of an individual, i.e. the number of other agents the individual interacts with. In this sense, it could also be seen as being part of the environment of an individual. However, considering that different individuals could choose to interact with different numbers of neighbors, I find it more suitable to discuss the parameter in this section. For the Prisoner's Dilemma on a one-dimensional lattice, the spread of cooperation depends on both the interaction and learning radii (ibid., p. 72). In small-world networks, whether the additional connections on the graph can block the expansion of regions governed by one strategy also depends on the interaction radius (ibid., pp. 78–79). Similar influences can be observed in the Stag Hunt. On one-dimensional lattices, the interaction radius influences how many stag hunters need to be present in a given region for this region to spread (ibid., p. 119). Which strategy ends up dominating on a two-dimensional lattice also depends on the interaction and learning radii (ibid., p. 125). For the Nash bargaining game, the interaction radius determines the outcome of the frontier competition on the one-dimensional lattice, i.e. the battle between different strategies at the border of competing regions of agents (ibid., pp. 164–166).

For some of the simulations described in the book, additional parameters have been added to further investigate the situations in which it proved to be difficult for the strategies reflecting the given moral intuitions to emerge. These parameters include the frequency of strategy updates, a discount rate for time-discounting, as well as the addition of especially influential agents named "leaders". The frequency of strategy updates specifies how often an individual is allowed to change his strategy according to one of the learning rules mentioned earlier. This parameter can have an impact on the Stag Hunt when it is played on a dynamic social network. As mentioned in the previous section, the interaction probabilities of the dynamic networks covered in the book quickly converge towards one and zero, effectively turning the dynamic network into a static one. For the Stag Hunt, this leads to the population rapidly dividing into one cluster for each strategy, whereby no agent can change his strategy anymore. In this case, the outcome of the simulation is practically determined by the initial distribution of strategies alone. If, however, the frequency of learning is decoupled from the frequency of interactions, the results can change. Depending on the probability of strategy updates, the majority of the network can turn

into Stag players (Alexander 2007, pp. 143–144). Another feature which can influence the Stag Hunt on dynamic networks is time-discounting. When employing a learning rule with a frequency lower than that of the interactions, we can choose to assign greater importance to more recent interactions compared to past ones. When adding a suitable discount rate to the simulations, the population can more reliably converge to a state in which every individual is playing Stag (ibid., pp. 144–147). These parameters can also impact the result of the ultimatum game on dynamic networks. Tweaking the learning frequency and discount rate can significantly increase the probability of the population being dominated by fair strategies (ibid., pp. 235–236). One additional component of interest is the emergence of correlated mutations as proposed by Peter Vanderschraaf and Alexander (Vanderschraaf and Alexander 2005). The idea here is to introduce so-called *leaders* to the simulation, which appear as random mutations among the population. If an agent turns into a leader, he is randomly assigned one of the available strategies, and the leader's neighbors subsequently imitate this strategy according to a predefined probability. Such an individual can thus change the strategies of multiple agents at once, hence called *correlated mutation*. When letting leaders emerge in the Stag Hunt played on a two-dimensional lattice, the population can rapidly converge to playing Stag (Alexander 2007, pp. 128–131). On bounded-degree networks, the same result can be observed when only allowing leaders to play Stag (ibid., p. 140).

What about the plausibility of differences in the parameters discussed in this section? The different learning rules could reflect differences in cognitive behaviour among the individuals. Insofar as the rules embody different levels of sophistication, they could also reflect differences in cognitive ability. Imitation of another person's strategy is a rather crude mechanism, while best response learning requires at least some basic counterfactual reasoning. The different forms of this parameter can therefore be seen as plausible. For the interaction and learning radii, the differences can also be justified rather easily. Different people have differently sized circles of acquaintances. In fact, the small-world networks used in some of the simulations essentially model exactly this property: Some individuals are better connected than the rest, they create the bridge edges across the ring of individuals. We can therefore assume that differences in these parameters are not problematic. Given the examples provided above, I therefore argue that differences in both the learning rules as well as in the interaction and learning radii are plausible and reliable.

For the three different additional features employed in some of the simulations, the situation is not as convincing. Intuitively, the different update frequencies and the existence of leaders seems realistic. They could be explained through different personalities governing a conservative or liberal approach to one's own beliefs, and differences in charisma among the individuals, respectively. Different forms of time-discounting could be plausible, as long as the variation is not too extreme. In general, however, the simulations conducted by Alexander only contain very few examples of these parameters in action. Additionally, in the cases where they have been used, they were brought up as possible ways to justify the emergence of our own moral intuitions (e.g. for Stag becoming the dominant strategy in the Stag Hunt played on dynamic networks). The majority of the simulations have not included these parameters. It is not clear whether the systematic addition of the features

would in fact lead to more differences in the outcomes of the models, or whether it would reduce such differences. The reliability criterion is thus not fulfilled. Therefore, I am going to ignore these parameters for the purposes of the argument about descriptive moral relativism.

4.4 Convergence

This section deals with leftover parts of the simulations which do not fit into one of the previous categories, but still influence the *convergence* of the model. This includes randomized mutation of the agents' strategies, but also the point in time at which the result of the model is evaluated according to some notion of stability.

In some simulations, agents can change their strategies purely based on chance. The dice are rolled each iteration of the model, and changes occur according to a predefined probability. These random mutations of strategies can influence the outcome of the simulations. We have already seen some influence of this in the previous sections. An important point to note is that mutations can both have a beneficial as well as an unfavorable impact on the stability of the results. For example, in the ultimatum game played on a small-world network, mutations cause the population to be reliably overtaken by unfair strategies (ibid., p. 231). In the Nash bargaining game, mutation can protect a population on a dynamic network from being overtaken by unequal splitting strategies (ibid., p. 198). On a two-dimensional lattice, mutations even let the population reliably converge to fair division (ibid., p. 175). Is it plausible that agents could randomly change their strategies without the influence of other relevant factors? In the biological interpretation of the evolutionary dynamics, this could easily be explained through actual mutation of the individuals' genes, leading to unpredictable changes in phenotypes and thus strategies of those individuals. In the cultural interpretation, however, the story is not quite as convincing. Alexander's justification is that people can experiment with new behaviours, and such experimentation can be seen as the cultural analogue of mutation (ibid., p. 19). A crucial difference between the two is that individuals presumably still experiment of their own volition. The changes in strategy would then not actually be random, but at least in part be determined by the cognition of the individual. In any case, though, mutations don't provide a convincing domain of relativization for the model. Like learning frequencies, time-discounting and leaders, mutation effects have not been systematically employed in all of the simulations covered in Alexander's book. Considering that the random effects can both enhance and impair the stability of the model, and thus can both increase and decrease the amount of resulting moral disagreements, the reliability criterion is not fulfilled. I am therefore also going to ignore this parameter for the purposes of the argument.

When running the simulations, the results need to be evaluated at some specific point in time. While this is being done by the modeller and not the model itself, it obviously has an influence on what is considered to be the result of the simulation. When using the replicator dynamics, the convergence of the model can be solved analytically, and there is thus no uncertainty about the further progression of the evolutionary dynamics.

For agent-based models, though, this is generally not possible. If random effects like mutations are used, the state of the population could, in theory, change at any point in the future. Alexander therefore uses a notion of the *dynamic stability* of a strategy, according to which a strategy is "unlikely to be driven out within a reasonably long time frame" (Alexander 2007, p. 280). He leaves this notion deliberately loose, though, and argues that it is sufficient for the purposes of his argument about the evolution of morality (ibid., pp. 280–281). When analyzing the domains of relativization of the model, it would be beneficial to have a clearer definition of stability which could be used to systematically evaluate the outcomes of the simulations. Considering that we need to rely on Alexander's reports, though, the notion of dynamic stability will have to do. I therefore assume that the results of the simulations described in the book can in fact be considered as final, and I am going to put aside any further potential issues with the stability criterion.

4.5 Results

In the last three sections, I have shown how different parameters of Alexander's model can lead to different strategies being selected through the evolutionary dynamics. There are many plausible changes in parameter values which reliably lead to different results for all of the four games discussed in the book. Concretely, I have identified five different parameters which fulfill the criteria of plausibility, reliability. These include the initial distribution of strategies, the topology of the social network, the payoff structure of the interpersonal decision problems, the learning rule used for updating strategies, as well as the expanse of an individual's neighborhood for interactions and learning. Various additional parameters in the category of *cognition* did not provide to be reliable enough causes for change, especially considering that they were only used sparsely in Alexander's experiments. Random mutation effects can both enhance and diminish the stability of the results, therefore also failing to fulfill the criterion of reliability. The five parameters for which the investigation was successful together (and each to some extent) have a considerable influence on the results of all of the four games covered in Alexander's model. I therefore argue that the substantiality criterion is also fulfilled. If the assumption made in section 3.3 holds (i.e. differences in dominant strategies for populations lead to differences in moral beliefs between those populations), then I have shown that the structural evolution of morality does indeed lead to descriptive moral relativism.

The five aforementioned parameters can now be considered as domains of relativization for an argument for metaethical moral relativism. When arguing for such a position, one could thus try to make it plausible that the best way to think about the truth status of moral judgements is that it is relative to one, multiple or all of these domains of relativization. While such an argument cannot be made in this paper, I am going to give a preliminary discussion about the metaethical implications of these results in the following chapter.

Chapter 5

Metaethical Discussion

Due to the limited scope of this paper, and because the structural evolution of morality leaves open a number of important questions which would need to be answered before one can properly engage in an argument about the truth status of moral norms, I cannot provide a conclusive argument for or against metaethical moral relativism. Nonetheless, I want to show what would have to be done in order to to make such an argument. Specifically, I want to present in this chapter what it takes to argue for metaethical moral objectivism (and thus against metaethical moral relativism) given the results about descriptive moral relativism.

Based on the results of the previous chapter, one could argue that the best way to think about the truth status of moral norms is that it is relative to those domains of relativization which we derived from Alexander's model. What is needed to defend moral objectivism against the stark moral diversity exemplified through the analysis of the previous chapter? Given the moral disagreements as part of the descriptive moral relativism, we need to argue that these disagreements can be resolved in some rational way. This way, we can justify that moral beliefs are still objectively true or false. There are a number of common objectivist responses to moral disagreements (Gowans 2019, sec. 5). An obvious one states that one set of moral beliefs is simply superior to all others. In the context of the structural evolution of morality, we could for example say that cooperation is better than defection. This way, people with a moral norm of defection would simply be wrong. But how do we justify this in the context of the interpersonal decision problems? After all, the evolutionary dynamics can lead to defection being the best strategy for some population of individuals. Why is one of the results superior to the other? One approach would be to take some notion of optimality which applies to one and not the other strategy profile. For example, we could take David Gauthier's approach and view Pareto-optimality as the decisive factor in what the right way to act in a given interpersonal decision problem is (see Gauthier 2013). Playing Cooperate in the Prisoner's Dilemma could then be seen as the objectively right thing to do, because the strategy profile of both players cooperating is Pareto-efficient. Of course we would also need to find such a notion for the other games, in order to resolve the leftover disagreements as well.

From this standpoint, there is little left that can be said. What about Alexander's

statements about the ethical implications of his theory? Let us remind ourselves of the quote at the end of *The Structural Evolution of Morality*:

> "Insofar as our moral beliefs provide solutions to interdependent decision problems, we cannot say that any one solution is better than any other — in an abstract sense — because, detached from our preferences, there is no absolute standard from which to judge." (Alexander 2007, p. 291)

There are two issues I take with the above statement. For one, the quote might make it look like a set of moral beliefs is specific to one person's explicit preferences. If this was true, then a moral norm could prescribe how to act in order to satisfy a specific preference, e.g. "act in such a way that maximizes the satisfaction of my preference of becoming a Rock star". This, however, misses the point. If two people are playing the Prisoner's Dilemma (not necessarily with each other), then the structure of the game already abstracts from the players' individual preferences. If the game didn't look like it does, in that mutual cooperation is the second-best outcome for both players, and mutual defection is the second-worst outcome, and so forth, then the two people would be playing a different game. We have to look at the situation from a higher level. A moral belief does not provide action guidance for choosing the strategy that is expected to satisfy my specific preference of, for example, becoming a Rock star. Rather, the norm prescribes the strategy which is expected to maximize the satisfaction of my preferences (given their ordered ranking) *in general*. Secondly, the part of the quote which states that the dependence on our preferences leaves no absolute standard from which to judge a given strategy is arguably moot. As I alluded to in section 4.2, by using evolutionary game theory to explain the emergence of these moral norms, we effectively already assume that the payoffs of a game are comparable among players. In the biological interpretation of the evolutionary dynamics, only nature needs to be able to make those comparisons, by virtue of counting the number of offspring or grandchildren an individual produces. In the cultural interpretation, and given the learning rules used in Alexander's model, the players themselves need to make those comparisons. If we view payoffs as measures of the satisfaction of one's preferences, then an objective standard from which to gauge differences in that satisfaction needs to be in place for the evolutionary dynamics to work out in the first place.

I think, however, that Alexander concedes at least the first of the two points I just made. In the video interview which already gave us some insight into Alexander's conception of morality in section 2.4, he formulates his view in a different way:

> "In my book *The Structural Evolution of Morality*, I defend one view as to how we are to think about moral theory. Each of us have our own particular preferences that we would like to realize in life, but the point is our preferences are not all compatible. Sometimes, what I want conflicts with what you want, and vice versa. Morality, by providing a set of principles in terms of how we ought to behave, provides the best way for each of us to satisfy our preferences

> to the greatest extent possible, given the constraints placed by other people." (Alexander 2015)

What about moral relativism, then? Alexander states that "morality becomes relative in that it becomes context-specific" (ibid.). This is a much weaker point than that of metaethical moral relativism. In fact, unless we take a strict deontological approach to morality, we can definitely reconcile at least some context-dependence with an objectivist moral theory. What is right to do depends on the specific situation we find ourselves in, i.e. the interpersonal decision problem we face, and the constraints the other players' preferences place on the satisfaction of our own preferences. The implications on the truth status of moral norms then ultimately hinge on how we think about the moral norms created by the structural evolution of morality.

Alexander's conception of morality, which I sketched out in section 2.4 as well as I could given the limited information available, leaves a lot of questions unanswered. This is not surprising, given that Alexander's theory is merely descriptive. It simply tries to provide an explanation of how our moral norms came to be. The relevant aspects to the metaethical discussion are thus largely missing. It is not clear what acting in a morally good way means in the context of the kind of morality brought up by the structural evolutionary processes. Based on Alexander's limited discussion about this, moral norms are supposed to reflect fast and frugal heuristics for maximizing expected utility in the long run, and these heuristics are turned into a set of emotional motivations through some psychological mechanism. Is the right thing to do in a given situation then to act according to those heuristics which have been programmed in our brains through the evolutionary processes? Or is it to act in such a way that maximizes the expected utility in the long run, independent of what the heuristics tell us? These are just two possible interpretations, but it is clear how they can conflict with each other. For example, consider that the evolutionary dynamics caused me (or my ancestors) to choose Cooperate in a Prisoner's Dilemma as the heuristic for utility maximization. At the time during which this heuristic was created, playing Cooperate might very well have been the optimal strategy in the long run. But now, I can find myself in environments where playing Defect is the better strategy. If I act according to my heuristic, I am going to get worse payoffs. If I act according to the expected utility maximization, I act against my adaptive heuristic which is supposed to reflect my moral belief. If acting according to one's instinctual, adaptive heuristics were considered to be morally good, then one could reasonably argue that moral truth is relative. It does then, after all, depend on the circumstances of the evolutionary processes described in chapter 4, which led to the creation of these heuristics. If, on the other hand, maximizing expected utility was morally good, then we could view exactly that as a common, objective moral norm. This norm can, under certain circumstances, lead to actions which we would instinctively call "fair" or "cooperative". But, it doesn't have to, and more often than not such actions could be considered morally wrong, because they fail to maximize expected utility in the long run. If we admit such an interpretation of the nature of morality as created by the structural evolution of morality, then we may just have to accept these circumstances for metaethical moral objectivism to hold.

Chapter 6

Conclusion

In this paper, I have argued that J. McKenzie Alexander's theory of the *Structural Evolution of Morality* leads to descriptive moral relativism. At the beginning of the paper, I have introduced the issue of autonomous moral reflection, which provided to be an important issue for the investigation into moral relativism. After a brief introduction into evolutionary game theory and Alexander's thesis and methodology, I tried to construct Alexander's conception of morality based on the structural evolution of morality. This sketch of a conception of morality helped to further the metaethical discussion at the end of the paper. I have reconstructed Alexander's argument about the implications of his theory on morality, and differentiated descriptive from metaethical moral relativism as two possible consequences of the structural evolution of morality. I then introduced the concept of the domain of relativization and used it as the basis for the methodology of the argument in this paper. The argument about descriptive moral relativism then started with a derivation of the potential domains of relativization from Alexander's model. I then analyzed the environmental and cognitive domains of relativization in Alexander's theory, as well as some additional convergence properties. The investigation lead me to argue that the structural evolution of morality does indeed lead to descriptive moral relativism. At the end of the paper, I discussed some of the metaethical implications of the descriptive moral relativism. The preliminary results showed that the consequences for ethics mostly depend on the understanding of morality that the structural evolution of morality brings forward. Further investigation into the nature of this morality is needed in order to provide a conclusive argument about metaethical moral relativism.

Bibliography

Alexander, J. McKenzie (2007). *The Structural Evolution of Morality*. Cambridge University Press. DOI: 10.1017/CBO9780511550997.

— (2015). *LSE Philosophy: J McKenzie Alexander*. URL: https://www.youtube.com/watch?v=hyWYGMg_GBI (visited on 06/21/2020).

— (2019). "Evolutionary Game Theory". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2019. Metaphysics Research Lab, Stanford University.

Axelrod, Robert (1984). *The Evolution of Cooperation*. Basic Books. ISBN: 0-465-02122-0.

Baghramian, Maria and J. Adam Carter (2019). "Relativism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University.

Binmore, Kenneth G. (1994). *Game Theory and the Social Contract: Vol. 1, Playing Fair*. MIT Press. ISBN: 9780262023634.

— (1998). *Game Theory and the Social Contract: Vol. 2, Just Playing*. MIT Press. ISBN: 9780262024440.

FitzPatrick, William (2016). "Morality and Evolutionary Biology". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2016. Metaphysics Research Lab, Stanford University.

Gauthier, David (2013). "Twenty-Five On". In: *Ethics* 123.4, pp. 601–624. DOI: 10.1086/670246.

Gowans, Chris (2019). "Moral Relativism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2019. Metaphysics Research Lab, Stanford University.

Grüne-Yanoff, Till (2011). "Evolutionary Game Theory, Interpersonal Comparisons and Natural Selection: A Dilemma". In: *Biology and Philosophy* 26.5, pp. 637–654. DOI: 10.1007/s10539-011-9273-3.

Haack, Susan (1996). "Reflections on Relativism: From Momentous Tautology to Seductive Contradiction". In: *Philosophical Perspectives* 10, pp. 297–315. DOI: 10.2307/2216249.

Hamilton, W. D. (1967). "Extraordinary Sex Ratios". In: *Science* 156.3774, pp. 477–488. DOI: 10.1126/science.156.3774.477.

Joyce, Richard (2005). *The Evolution of Morality*. MIT Press. ISBN: 9780262101127.

Nagel, Thomas (2012). "Ethics without Biology". In: *Mortal Questions*. Canto Classics. Cambridge University Press, pp. 142–146. DOI: 10.1017/CBO9781107341050.012.

Shermer, Michael (2004). *The Science of Good and Evil*. Henry Holt and Company. ISBN: 9781429996754.

Skyrms, Brian (2014). *Evolution of the Social Contract*. 2nd ed. Cambridge University Press. DOI: 10.1017/CBO9781139924825.

Skyrms, Brian and Robin Pemantle (2000). "A dynamic model of social network formation". In: *Proceedings of the National Academy of Sciences* 97.16, pp. 9340–9346. DOI: 10.1073/pnas.97.16.9340.

Vanderschraaf, Peter and J. McKenzie Alexander (2005). "Follow the Leader: Local Interactions with Influence Neighborhoods". In: *Philosophy of Science* 72.1, pp. 86–113. DOI: 10.1086/428077.

Weibull, Jörgen W. (1995). *Evolutionary Game Theory*. MIT Press. ISBN: 9780262231817.

Declaration of Authorship

I hereby declare that I have composed the presented paper independently on my own and without any other resources than the ones indicated. All thoughts taken directly or indirectly from external sources are properly denoted as such.

Munich, June 21st 2020