Reason, Evolution, and the Possibility of Cooperation

by

Katharine Browne

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Philosophy
University of Toronto

© Copyright by Katharine Browne 2012
Reason, Evolution, and the Possibility of Cooperation

Katharine Browne
PhD 2012
Department of Philosophy
University of Toronto

Abstract

“Cooperation” has distinct meanings in biological and moral contexts. In nature, “cooperation” is commonly equated with “altruism,” and involves an apparent fitness cost to the actor. In the moral context, it is often employed to describe the behaviour required by a particular subset of morality, namely that of distributive justice.

The existence of cooperation in nature poses a difficulty for those who seek explain it in evolutionary terms. There is an analogous problem in normative moral theory of reconciling cooperative behaviour with rationality. The constraints imposed by natural selection in the former context and by rationality in the latter make explaining or justifying cooperation difficult.

Insofar as the social contract tradition is concerned with articulating or justifying the terms of social cooperation, these two problems are united through the contract metaphor. I examine these two structurally similar problems through the lens of the social contract tradition.

In the descriptive arena, I argue that cultural group selection provides the most plausible explanation of the emergence of altruistic behaviours in nature. In the
normative context, I argue that David Gauthier’s argument for the rationality of adopting the disposition of constrained maximization provides a defensible route to reconciling morality with rationality.

I draw two conclusions with respect to how these two enterprises are connected. First, I argue, contrary to many critics of an empirically informed ethics, that the descriptive and normative projects are very much dependent upon one another. Insofar as culture is required for a descriptive account of the emergence of cooperation, and to the extent that reasoning about which norms ought to govern our interactions plays a role in their transmission, our descriptive account not only leaves room for normative considerations, but in fact requires them.

Second, I argue that there is a convergence in the outcomes of both the descriptive and normative projects. I show that the explanation of the existence of cooperation that I favor also provides us with an explanation of the emergence of dispositions that structurally resemble those that Gauthier defends as rational. And thus we arrive at an account that brings together rationality, evolution, and morality.
Acknowledgments

It is with a tremendous sense of relief that I write these acknowledgements. This is a project whose future looked bleak not long ago, and were it not for the heroic efforts of Paul Thompson and the wonderful committee that he helped to assemble, this thesis would not be what it is today.

I thank Paul for breathing new confidence into this project and making it possible to carry through to completion. Much gratitude is also owed to my committee members, Wayne Sumner and Denis Walsh, for their acute comments and kind encouragement throughout this process. Ours was a brief assembly, but it is one for which I am immensely thankful.

Thank you to my mother and her sharp eyes and proofreading efforts of the penultimate version of this thesis. And to my father, the first philosopher I have known. I am most fortunate to have had his unwavering support and guidance over the course of my education. My gratitude to him is immeasurable. It is fitting to say here that philosophy is in the genes. But it is also true that people imitate those whom they admire. My mother told me that I when I was much younger I would put my father’s shoes on, pick up his briefcase, and pretend to go off to work. As I close this chapter and look onto the next I am very grateful for the footsteps that I have in which to follow.
# Contents

Abstract .......................................................................................................................... ii  
Acknowledgements ................................................................................................. iv  
Introduction. Two Problems of Cooperation ......................................................... 1  
Chapter 1. Evolutionary Game Theory and The Social Contract ....................... 10  
Chapter 2. The Folk Theorem and Cooperation in Repeated Games ............... 36  
Chapter 3. Biological Group Selection ................................................................. 56  
Chapter 4. The Significance of Culture ................................................................. 83  
Chapter 5. Normative Social Contract Theory  
    and David Gauthier’s Contractarianism ......................................................... 104  
Chapter 6. Reconciling Reason, Evolution, and Morality ................................... 141  
Bibliography ............................................................................................................ 157
INTRODUCTION

Two Problems of Cooperation

Nature is rife with examples of cooperative behaviours. From honeybee societies, altruistic vampire bats, to the enormously complex civilizations formed by human beings, cooperation lies at the foundation of all social interactions.

Finding examples of cooperation is not difficult. Explaining how it is possible, however, is less straightforward. The existence of cooperation in nature (or altruism, as it is most commonly referred to in this context\(^1\)) is a bit of a mystery, in light of evolutionary theory. The reason is that altruism seems to involve a fitness cost to its actor, a characteristic that is difficult to make sense of given that nature selects against fitness-decreasing traits.

\(^1\) I will use the terms “cooperation” and “altruism” interchangeably and in a very general way to refer to behaviour that runs contrary to individual interest, which can either be construed in terms of fitness in the biological context or utility-maximization in the normative context. There are differences between them, but none is significant for my purposes here, which is to give an account of what I take to be the most likely explanation and justification of these behaviours. It is not unusual to see these equivalences drawn in the literature. Sober and Wilson say that, “prevalent among game theorists is their use of the word cooperation rather than altruism… the word cooperation is used by evolutionary game theorists, presumably because it is easier to think of cooperation as a form of self-interest. The behavior is the same but it is labeled differently.” Elliott Sober and David Sloan Wilson, Unto Others: The Evolution and Psychology of Unselfish Behavior (Cambridge, MA: 1998), 84. And Richerson and Boyd point out that they “use the word cooperation to mean costly behavior performed by one individual that increases the payoff of others. This usage is typical in game theory, and common, but by no means universal in evolutionary biology. It contrasts with ordinary usage in which cooperation refers to any coordinated, mutually beneficial behavior.” Robert Boyd and Peter J. Richerson, “Culture and the Evolution of the Human Social Instincts,” in Roots of Human Sociality, eds. N.J. Enfield and Stephen C. Levinson (New York: Berg Publishers, 2006), 454.
There is an analogous problem in normative moral theory. Rationality is most commonly understood to require that agents maximize their utility. And yet morality calls for a constraint on self-interested pursuits, a constraint that is hard to square with a maximizing conception of rationality.

One way in which these two problems are related is through the social contract. For the social contract theorist, morality is constituted by a set of principles that individuals agree to. These principles govern social interaction, and become necessary when individuals need to band together in cooperative ways. Morality comes into existence out of the need for cooperation, and sets the terms of it. Thus, insofar as the social contract tradition is concerned with articulating or justifying the terms of social cooperation, these two problems are united through the contract metaphor.

Within the contract tradition there is a descriptive branch and a normative branch. Descriptive approaches describe the origin of the social contract and seek to explain how cooperation occurs. The first problem of cooperation described above falls to this branch. This is the tradition of Rousseau and Hume. There has been a recent surge of interest in approaching the explanation of the emergence of the social contract in terms of evolutionary processes. Drawing from (comparatively recent) findings in evolutionary biology and anthropology, this “naturalistic” approach to the social contract promises to shed light on the underlying mechanisms responsible for

---

2 While he is not a social contract theorist, I include Hume here insofar as he is concerned with explaining the origin of cooperative norms in a way congenial to that theory. David Gauthier articulates the contractarian character of Hume’s theory of justice and convention in “David Hume, Contractarian,” *The Philosophical Review* 88, no. 1 (January 1979): 3-38.
the origins and sustenance of human sociality. This line is pursued in contemporary literature by Brian Skyrms, Ken Binmore, Elliott Sober and David Sloan Wilson, and Peter Richerson and Robert Boyd. There is no shortage of theories to explain cooperation: kin selection, reciprocal altruism, indirect reciprocity, replicator dynamics, biological group selection, and cultural group selection have all been invoked, for example. There is, however, no agreement on which is the appropriate one.

Normative approaches aim to justify the terms of the contract. The second problem of cooperation identified above falls to this tradition. Successful cooperation requires constraint on the pursuit of individual self-interest. The social contract justifies the constraints characteristic of moral and political obligation in the fact that they would be required by principles that would be selected by the individuals who are constrained by them. Agreement to these principles is grounded in reason. Just as there is no shortage of dispute between alternative explanations in the descriptive camp, so too is there controversy among candidate normative theories.

While normative social contract theories share the view that justice and social cooperation have a basis in reason, they differ according to which conception of reason is employed. Stephen Darwall distinguishes two kinds of social contract

---

3 The latter two pairs do not explicitly identify themselves as social contract theorists but are concerned with the same issues surrounding cooperation as those that do.

4 As we will see, the disagreement doesn’t end here. Even if we settle the question of which of contractarianism or contractualism to endorse, there are further disputes within each subgroup, regarding, for example, what are the terms of cooperation that would be selected through bargaining, how to (or whether one can) guarantee that rational individuals will comply with those terms, and so on.
theories. “Contractarians,” represented by Thomas Hobbes and David Gauthier, appeal to “rationality” in the sense of utility-maximization to justify the principles of morality. According to this view, individuals agree to certain principles because it is in their interest to do so. “Contractualists,” represented by (among others) John Rawls and Thomas M. Scanlon, by contrast, appeal to the notion of “reasonableness,” according to which principles are chosen on the basis of morally relevant interests or values, which are established through argument. Which account we choose has significant implications. If reason is understood as rationality where this means the pursuit of one’s self-interest (as it is according to contractarianism), and morality requires that we curb self-interested pursuits, there is an apparent tension between reason and morality. On the other hand, if reason is understood as reasonableness (as it is in contractualism) then we escape the seeming incompatibility with reason and morality, but are left with a less basic conception of reason, and this requires defense.

In this thesis I will do three things: (1) defend the theory of cultural group selection as the most plausible explanation for widespread human cooperation, (2) defend David Gauthier’s contractarian argument for the rationality of developing a disposition to moral behaviour, or to “constrained maximization,” and (3) articulate what the relationship is between the descriptive and normative enterprises. This thesis is organized according to these aims.

The first four chapters take up the first aim, and constitute an examination of the descriptive problem of cooperation and prominent attempts to explain it. Taken as

---

requiring an explanation is the fact that human beings engage in high levels of cooperation between non-related individuals and display (what appears to be) genuinely altruistic tendencies towards others. From a purely biological standpoint, this behaviour is mysterious, primarily because of the scope and scale of these interactions. Two standard mechanisms invoked to explain cases of altruistic behaviours are kin selection and reciprocal altruism.\(^6\) Parents, for example, act in such a way that appears to directly reduce their own fitness and promote the fitness of their offspring. This invites the question how such behaviour could arise. Kin selection gives us an answer. Since offspring contain on average one half of the genes carried by either parent, we see that caring for offspring, which is itself costly to the individual, is a behaviour that ultimately promotes the survival of one’s genes. Reciprocal altruism permits the same kind of explanation in cases where individuals are not related. Nitpicking in birds, grooming among chimpanzees, and sharing of blood among vampire bats are all examples of apparently altruistic behaviour. On closer inspection, however, we see that the apparent fitness costs imposed by these acts are recouped through reciprocity and can thus be made sense of within an evolutionary framework.

While much altruistic behaviour in both human and non-human animals can be explained through the evolutionary mechanisms of kin selection and reciprocal altruism, these mechanisms are not sufficient to explain all of the cooperation we

find. In particular, they do not explain cooperation between non-related human beings, and between individuals who have no expectations of meeting each other again. Consequently, various other mechanisms have been proposed to explain how what Richerson and Boyd refer to as “ultra sociality” might have emerged. These include evolutionary game theory (as we will see in Skyrms, chapter one), more sophisticated models of reciprocity (as we will see in Binmore, chapter two), biological group selection (as we will see in Sober and Wilson, chapter three), and cultural group selection (as we will see in Richerson and Boyd, chapter four). I will examine these approaches with an eye to evaluating their adequacy in accounting for all relevant facets of human cooperation. I will argue that the accounts of Skyrms, Binmore, and Sober and Wilson, all fall short of the mark; that is, they fail to provide adequate accounts of the emergence of widespread cooperation. I will argue that cultural group selection, by contrast, explains what the other theories cannot and does so in a way that is compatible with evolutionary processes. Thus I will conclude that cultural group selection emerges as the most plausible candidate for explaining the emergence of cooperation in humans.

In chapter five I will turn to the structurally similar normative project of rationally justifying cooperation. Of central focus in this chapter is David Gauthier’s defense of the rationality of disposing oneself to cooperative behaviour. Gauthier endorses an optimizing rather than maximizing conception of rationality, which
promises to render moral (or cooperative)\textsuperscript{7} behaviour compatible with rational behaviour. I will argue that this reconceptualization of rationality is warranted and that, consequently, we are able to show that the disposition to what Gauthier refers to as “constrained maximization,” which calls for identifiably moral action, is also advantageous and thus rational. Thus I will conclude that rationality and morality can be reconciled.

What we take to be the import of the conclusions drawn from the descriptive and normative sides will depend in part on how one regards the relationship between the descriptive and normative traditions. Drawing out the implications of the first two sections of the thesis is the task of chapter six. There are two issues in particular that I want to address here. The first concerns what relevance the traditions have to one another.

The descriptive and normative projects can be taken as separate endeavours or as complementary or related in some way. Haunted by the unbridged “is/ought” distinction, the notion of an empirically informed contract is sometimes approached with caution by some moral philosophers. These critics contend that normative guidance about what we ought to do cannot be derived from facts about nature alone; that is, that the descriptive cannot inform the normative. The worry is thus that taking seriously an evolutionary approach to the social contract is liable to blend the

\textsuperscript{7} “Cooperation” is perhaps most commonly employed to refer behavior required by the subset of morality concerned with distributive justice. One of my primary aims in this thesis is to show that this behavior can be reconciled with rationality. I use “morality” and “cooperation” interchangeably for ease of use, but this should not be taken to imply that in claiming that cooperative behavior can be reconciled with rationality I thereby claim that all of moral behavior is rational. I leave that an open question, beyond the scope of my aims here.
descriptive and normative in unjustifiable ways by trying to derive normative conclusions from a descriptive account. This, however, is not the aim of most prominent figures in the field. One of the motivations behind the recent naturalistic turn is indeed to eliminate the normative. Skyrms and Binmore claim that an adequate descriptive account will render normative considerations unnecessary. As Binmore puts it in his *Natural Justice*: “If morality evolved along with the human race, then asking how we ought to live makes as much sense as asking what animals ought to exist, or which language we ought to speak.”

Against this view, I will argue that the descriptive and normative projects are very much dependent upon one another, and that we can maintain this position without making any unwarranted leaps across the is/ought barrier. As I will argue in chapter four, cultural group selection provides the most plausible explanation of the emergence of cooperation in nature. Invoking culture in the descriptive account, I will contend, opens the door to normative considerations. More specifically, I will argue that cultural transmission of norms depends in part on whether others adopt those norms. Whether others adopt those norms depends on whether there are what individuals take as good reasons to adopt some norms and not others. To do so is to engage in normative deliberation. Thus, insofar as culture is required for a descriptive account of the emergence of cooperation, and to the extent that reasoning about which norms ought to govern our interactions plays a role in their transmission, our

---

descriptive account not only leaves room for normative considerations, but in fact requires them.

The second implication concerns the relationship between the descriptive and normative accounts I have defended. Here I will contend that these accounts converge. More specifically, I will contend that the explanation of the existence of cooperation that I defend coincides with the emergence of dispositions that resemble those that Gauthier defends as rational. Cooperation (as I will argue in chapter four) requires prosocial dispositions that call for constraint on the pursuit of self-interest. Rationality (as I will argue in chapter five) calls for the formation of the disposition to constrained maximization, which likewise requires similar constraints on the pursuit of self-interest. Thus, insofar as the constraints of constrained maximization are characteristic of morality, we arrive at an account that brings together evolution, rationality, and morality.
CHAPTER ONE

Evolutionary Game Theory and The Social Contract

Cooperation calls for a constraint on self-interested pursuits. Successful cooperation generates a mutually advantageous outcome. It also requires that enough individuals participate. If the terms of cooperation are not agreeable to all agents, then cooperation is not likely to be successful. One central aim of the social contract is to articulate those terms.

Classical (or rational choice-based) game theory (RGT) provides us with an analysis of the interactions of ideally rational agents, and has been applied to the social contract to handle two key problems of the social contract, namely articulating the terms of cooperation and ensuring compliance with those terms. RGT relies on the *homo economicus* model of man, according to which individuals are presumed to be self-regarding utility-maximizers. As a descriptive tool, it is intended as a best approximation of how real-life agents behave, and can be used to explain and predict the outcomes of various interactions. As a normative tool, RGT is used to prescribe which actions rational actors (understood as utility-maximizers) ought to take.

From a game theoretic perspective, the terms of cooperation constitute a Nash equilibrium, which is a set of strategies where no individual has an incentive to
change his strategy, given the strategy employed by his opponent.¹ Often there will be multiple equilibria for any one particular game. Articulating the terms of cooperation will therefore involve solving what is known as an “equilibrium selection” or “coordination” problem. Once those terms are set there is a second issue of how to guarantee compliance with those terms. Making an agreement is one thing; keeping it another, a problem that becomes quite acute given that cooperation is not always in equilibrium and the possibility of sophisticated free rider strategies.

As we will see, RGT has limitations, both in its descriptive and normative powers. As a consequence, some game theorists have moved to evolutionary processes in an attempt to resolve what RGT cannot.² My aim in this chapter will be to examine the evolutionary game theoretic approach (EGT) employed by Brian Skyrms³ to solve the problems of the social contract. I will argue that EGT approaches like Skyrms’s fail because they do not capture the high level cognitive functions (viz., rationality) of human beings. These functions permit high levels of cooperation but also leave room for the possibility of free rider strategies. EGT approaches do not account for this and, thus, fall short in accounting for the widespread cooperation that we seek to explain.

¹ For a technical treatment of this concept, see Jorgen Weibull, Evolutionary Game Theory, paperback ed. (Cambridge, MA: MIT Press, 1997), 14-17.


Game Theory Applications and Limitations

Game theory is the study of strategic interaction between players in a defined game. Each player in the game is assumed to be rational, meaning that he or she is able to identify and rank preferred outcomes, is able to identify a means to that preferred outcome (defined in terms of utility), and is able to select an action that will lead to that outcome based on his or her knowledge or anticipation of what the other player(s) will do.4

The shift to EGT is primarily a result of the shortcomings of RGT in resolving two key problems of the social contract, namely coordination and cooperation problems.5 A coordination problem (or equilibrium selection problem) is one where there are multiple equilibria and individuals aim to coordinate on one of them. The “Divide the Cake” game6 is one such example. It runs as follows. Two people are dividing a cake. Each writes down the percentage he or she wishes to claim. If the total of the two claims is greater than 100%, then neither person gets any of the cake. If the claims are equal to or less than 100%, each player will get his or her claimed share. In this case, a demand for a portion of the cake is a Nash equilibrium if no player can do better by demanding more, given what the other player has demanded. Thus, all sets of claims that sum to 100% are Nash Equilibria. However, as Robert

---

4 For an elegant account of the central elements of classical game theory, see Sugden, “The Evolutionary Turn in Game Theory,” 115.

5 Coordination problems are characterized by the presence of multiple equilibria. Resolution to these problems involves selecting (or “coordinating” on) a unique equilibrium. Cooperation problems like the Prisoner’s Dilemma, by contrast, involve out-of-equilibrium play.

6 Skyrms employs this game as his chief model of the equilibrium selection problem in Evolution of the Social Contract.
Sugden says: “The Holy Grail is a solution concept which, for every game, picks out one and only one combination of strategies as the solution.”\(^7\) This standard is not met in the Cake problem and, thus, game theory does not give a satisfactory solution.

Furthermore, in experimental settings, when subjects who are equally situated\(^8\) are asked to divide the cake, there is an overwhelming consensus that the cake should be divided evenly. What this suggests is that individuals are employing a particular norm of fair division—viz., that a fair division in such games is a 50/50 split.\(^9\) But why this is the case is not explainable in terms of RGT alone. From the perspective of RGT, there is nothing special about the 50/50 division that would explain why it is the case that individuals are inclined to split the cake evenly. Rather, RGT will be indifferent between any division that sums to but does not exceed 100% of the cake. All that RGT can explain, for example, is why it is the case that Player 1 demands 40% of the cake when Player 2 demands 60%. In this case, demanding 40% is Player 1’s unique best response to Player 2’s demand. The same goes for Player 1’s demand for 25% in response to Player 2’s demand for 75%, and so on.

Coordination problems lie in contrast to cooperation problems. The Prisoner’s Dilemma provides a formalization of the problem of cooperation. It involves two accomplices who are caught for committing a crime, interrogated separately, and

---

7 Sugden, “The Evolutionary Turn in Game Theory,” 115.

8 Meaning that there is no reason that might give an individual a claim to a greater portion of the cake than the other (e.g., he or she baked the cake, or is hungriest, or is thinnest, or is fattest, or has greater bargaining strength, etc.). Asymmetric bargaining problems, by contrast, require proportional solutions, which can be determined by any number of different axiomatic approaches (e.g., Nash, Kalai-Smorodinsky, etc.) In general, there is no widely accepted solution to such problems.

offered a deal. If one player incriminates the other, or “defects,” while the second remains silent, or “cooperates,” he will be given a sentence of one year, while the other player gets four. If both remain silent, both will be sentenced to two years, but if both defect, both will receive three years. The following matrix represents this game.

<table>
<thead>
<tr>
<th></th>
<th>Cooperate</th>
<th>Defect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperate</td>
<td>2 years, 2 years</td>
<td>4 years, 1 year</td>
</tr>
<tr>
<td>Defect</td>
<td>1 year, 4 year</td>
<td>3 years, 3 years</td>
</tr>
</tbody>
</table>

The first number of each pair represents Prisoner 1’s possible outcomes; the second number, Prisoner 2’s. In this particular case, no matter what the other player does, defecting is the utility-maximizing response. However, given that each player is rational, both will employ this equilibrium strategy, which will lead to a situation that is less preferred to the one where both cooperate. Rationality thus sometimes leads

---

10 There is an important distinction to note between one-shot Prisoner’s Dilemmas and iterated versions. As we will see, the equilibrium structure changes when certain games like the Prisoner’s Dilemma are repeated, and cooperative equilibria become available.
players to a suboptimal outcome. Furthermore, experimental settings reveal that subjects often cooperate at much higher levels than RGT would predict.\textsuperscript{11}

Equilibrium selection problems like the Divide the Cake game and cooperation problems like the Prisoner’s Dilemma challenge not only the empirical adequacy of RGT but also its prescriptive power. RGT is unable to provide an explanation of observed behaviour in Divide the Cake games and Prisoner’s Dilemmas. It also recommends actions that lead to suboptimal outcomes. Insofar as equilibrium selection and Prisoner’s Dilemmas are central to the social contract, and insofar as RGT fails to solve them, this suggests that RGT alone is not sufficient to address the problems of the social contract.\textsuperscript{12}

**Naturalizing the Social Contract**

One central aim of naturalists who tackle the social contract is to explain how individuals solve equilibrium selection and cooperation problems. More specifically, given the observation that we are able to coordinate on particular equilibria out of

---


\textsuperscript{12} Sugden says that: “According to folk history, the evolutionary turn in game theory was a response to the intractability of these two problems within the classical framework.” (“The Evolutionary Turn in Game Theory,” 220). Recognition of these limitations of classical game theory has led the field in two distinct directions. The first is known as the “refinement program.” Those who engaged in this program sought to make refinements to the Nash equilibrium concept by tinkering with the definition of rationality employed to rule out some Nash equilibria as insufficiently rational according to whatever definition of rationality was being employed. The problem with this methodology was that, after time, so many definitions of rationality had been proposed that any Nash equilibria could be justified by at least one definition of rationality. The second approach, which will be our focus here, appeals to evolutionary game theory to tackle the problems classical game theory leaves unresolved.
many, and given that we do cooperate (neither of which can be explained by RGT alone), the question then becomes how best to explain this. Both Skyrms (as we will see in this chapter) and Binmore (to whom we will turn in the next chapter) look to evolution to resolve equilibrium selection problems. Binmore and Skyrms both maintain that the particular social contract of a society (or its terms of cooperation) is in place because that is where evolution led it. In their view, norms of fairness are adaptations that permit individuals to coordinate on a particular equilibrium, which then becomes a part of a society’s social contract.

My primary concern in this thesis is with cooperation problems, and so I will only provide a brief sketch of their accounts of equilibrium selection in order to outline the general methodology employed by them. Since the Prisoner’s Dilemma is often regarded as the archetypical model for cooperation, that will be of particular interest. As Harms and Skyrms say:

Social beings can perform many acts which cost them little and benefit others more, like warning of danger, sharing excess food, or merely leaving another’s property alone. An act which costs one unit of fitness to perform, and benefits the recipient two, generates a Prisoner’s Dilemma. As such, the Prisoner’s Dilemma has come to be a theoretical microcosm for the study of the stability of cooperation and its benefits and thus for the evolution of moral

---

13 We will see in the next chapter that Binmore retains RGT in the analysis of cooperation.

14 Economists of Binmore’s stripe think that the one-shot Prisoner’s Dilemma is a misguided model for cooperation; that is, that if interaction did take place in this way, we would never have evolved as cooperative creatures. See, for example, Binmore, Natural Justice, 63. Philosophers, on the other hand, seem much more interested in proposed solutions to the Prisoner’s Dilemma, but these solutions ultimately transform the game from a cooperation problem to one where the cooperative outcome is also an equilibrium of the game.
behavior. Since cooperation is so commonly observed in nature, the theoretical challenge is to come up with solutions to the Prisoner’s Dilemma, plausible modifications of the situation such that cooperation can stabilize.\textsuperscript{15}

As we will see in this chapter and in the next, there are various ways that have been proposed to resolve the Dilemma and achieve cooperation. In particular, by repeating the game we will see a transformation of the equilibrium structure such that cooperation emerges as possible.

Skyrms’s central aim is to show that EGT is more able than RGT to explain the evolution of the social contract. He says: “My aims are explanatory rather than normative, and one of the themes of this book [Evolution of the Social Contract] is the superior explanatory power of evolutionary game theory when compared with game theory based on rational choice.”\textsuperscript{16}

Skyrms addresses the equilibrium selection problem through the Divide the Cake problem, and analyzes the problem of cooperation through what is known as the Stag Hunt game. According to Skyrms, the Prisoner’s Dilemma can be transformed into a Stag Hunt when the game is repeated,\textsuperscript{17} and is better able to model key features of cooperation than the one-shot Prisoner’s Dilemma. In what follows I will examine


\textsuperscript{16} Brian Skyrms, “Précis of Evolution of the Social Contract,” Philosophy and Phenomenological Research 59, no. 1 (March 1999): 217. Skyrms and Binmore have practical applications in mind when they seek evolutionary explanations of norms. Specifically, they want to use their findings about the source of our norms to make efficient reforms to our current social contract. According to them, any kind of reforms that we recommend must be within what it is feasible for people to do. Thus it is important to uncover the mechanisms underlying our cooperative tendencies.

\textsuperscript{17} Skyrms, The Stag Hunt and the Evolution of Social Structure, 4-6.
Skyrms’s analysis of the Divide the Cake and Stag Hunt games and will argue that EGT fails to provide an adequate solution to either. Before addressing the specifics of each game, let us turn to the central features of EGT.

**Evolutionary Game Theory**

RGT takes individual agents as self-interested utility-maximizing creatures. The success of various strategies is measured in terms of how much utility they produce for the ones who employ them. In the case of EGT, by contrast, success is measured in terms of fitness. Replicator dynamics\(^{18}\) is a tool used to analyze the fitness of particular strategies, according to how quickly they spread throughout a population. This approach looks at a population composed of at least two different strategies and determines which strategies will do better, in terms of fitness, than others. This will be determined by which strategies yield a greater payoff when interacting with other members of the population. And if one strategy is fitter than the other, that strategy will spread and begin to take over the population.

In short, replicator dynamics is a way to model the emergence of various strategies within a population. Two key elements used in analyzing how populations settle on one particular strategy or set of strategies are the evolutionary stability of the strategy (ESS)\(^ {19}\) in question, and size of its basin (or zone) of attraction.\(^ {20}\)

---

\(^{18}\) Skyrms employs this as his chief tool in *Evolution of the Social Contract* and *The Stag Hunt and the Evolution of Social Structure*.

Evolutionary stability is judged in terms of a strategy’s ability to resist mutant invasions. A strategy is said to be evolutionarily stable if, once a population is composed of individuals who mostly employ that strategy, that population is uninvadable by mutant strategies. That is, any invading strategy will do strictly worse when paired against the host population strategy than the host strategy will do when paired against itself and, thus, will not reproduce (and will eventually die off) in that population. The size of a strategy’s basin of attraction indicates how likely it is that any population will settle into a particular strategy. Given any number of different possible initial population structures, a variety of different strategies can emerge. A strategy has a large basin of attraction if it has a tendency to emerge from a large set of initial conditions, and small if it does not. According to Skyrms’s evolutionary account, the solutions that are arrived at in equilibrium selection and cooperation problems are not the products of rational choice; we are instead led to them by evolution.

It is important to note here the similarity between the equilibrium concepts of both EGT and RGT approaches. Skyrms points out that:

In Evolutionary theory, as in classical game theory, there is strategic interaction in the form of frequency-dependent selection, but there is no presumption of rationality—let alone common knowledge of rationality—operative. Indeed the organisms that are evolving may not even be making central elements of ESS in *Game Theory Evolving*, 2nd ed. (Princeton, NJ: Princeton University Press, 2009), 229-246.

20 Weibull, *Evolutionary Game Theory*, 75.
decisions at all. Nevertheless, game theoretic ideas have been fruitfully applied to strategic interaction in evolution. And the key equilibrium concept is almost the same.\textsuperscript{21}

Skyrms goes on to say that we can see a similarity between evolutionary stability and the Nash equilibrium concept:

The most striking fact about the relationship between evolutionary game theory and economic game theory is that, at the most basic level, a theory built of hyper-rational actors and a theory built of possibly non-rational actors are in fundamental agreement….Criticism of game theory based on the failure of rationality assumptions must be reconsidered from the viewpoint of adaptive processes. There are many roads to the Nash equilibrium concept, only one of which is based on highly idealized rationality assumptions.\textsuperscript{22}

EGT does, however, yield results that sometimes differ from RGT. In particular (as we will see), it narrows the range of possible equilibria available in situations, like the Divide the Cake game, where RGT remains indifferent between them. It also promises to explain behaviours like cooperation, which, as we have seen, often run contrary to self-interest. In Skyrms’s own words:

\begin{quote}
[A]s we look more closely at the theory in the evolutionary and rational choice settings, differences begin to emerge….Standard evolutionary
\end{quote}


\textsuperscript{22} Skyrms, “Game Theory, Rationality and Evolution of the Social Contract,” 273.
dynamics, the replicator dynamics, does not guarantee elimination of weakly dominated strategies. In a closely related phenomenon, when we consider extensive form games, evolutionary dynamics need not eliminate strategies which fail the test of sequential rationality. Going further, we shall see that if we generalize the theories to allow for correlation, we find that the two theories can diverge dramatically. Correlated evolutionary game theory can even allow for the fixation of strongly dominated strategies. These are strategies which fail under even the weakest theory of rational choice—the theory that players are in fact Bayes rational.23

We can now turn to see how Skyrms employs replicator dynamics to analyze the Divide the Cake and Stag Hunt games.

**Fairness and Divide the Cake**

Skyrms tries to demonstrate that in the Divide the Cake game demand half is the only pure strategy24 that is an ESS. According to him, we intuitively think that we ought to each demand 50% because doing so is an evolutionarily advantageous strategy. In other words, Skyrms thinks that individuals who demanded 50% tended to do better than those who did not, and thus the strategy of demanding 50% evolved.

Rather than cashing out payoffs of this game in terms of pieces of cake, payoffs are understood in terms of Darwinian Fitness—i.e., number of offspring. Individuals who receive a greater share of the cake thus translates to those individuals

---


24 Pure strategies lie in contrast to mixed strategies. An agent who employs the same strategy (e.g., “demand fair”) on every play employs a pure strategy. An agent who employs a mixture of strategies (e.g., “demand fair 60% of the time; play greedy 40% of the time”) plays a mixed strategy.
producing a greater number of offspring than those who receive less. If, for example, demanding fair (i.e., 50%) generally results in an individual receiving more cake than one who makes a greedy demand, then, on the evolutionary model, that individual will produce a greater number of offspring. The proportion of individuals in a population who demand fair will thus increase.

The frequency of encountering a particular strategy will depend on the initial concentration of that strategy in the population. When there are only pure strategies, fair (viz., demand 50% of the cake) will soon come to dominate the population. To illustrate this, suppose the members of a population have all settled into a strategy of claiming half the resource. If some newcomers arrive who claim more than half, they will get nothing, for the sum of their claim and that of whomever they are paired with will exceed 100%. If other newcomers arrive who claim less than half, they will get less than half. So a mutant invasion of a strategy other than claiming half does strictly worse than claiming half does against itself, and as a result the mutants will tend toward extinction.

On the other hand, if all in a population claim any share other than half, an incoming mutant group of individuals who claim half will do better on average and drive the original strategy to extinction. Hence claiming half is an evolutionarily stable strategy. This would seem to point to a general explanation of the fair division norm: evolution eventually carries any population to this norm.25

25 Skyrms, *Evolution of the Social Contract*, 9-11. Skyrms argues that the strategy demand 100% is also an equilibrium, but an unstable one, as it can be invaded by a modest proportion of mutants who demand less than 50%. For while these mutants will get nothing when paired with those who demand 100%, they will do better than those demanding 100% when they are paired with each other.
The above applies to cases where one single strategy takes over a population. But Skyrms notices the possibility of the emergence of what he calls “polymorphic states,” where, within a population, different proportions of that population are playing different strategies.\(^\text{26}\) It is possible to get to a state where, say, 75% of the population is greedy and demands 80%, and 25% of the population plays modest and demands 20%. In this case, when greedy meets modest, he gets 80%, but when he meets another greedy he gets nothing. Modest, on the other hand, can expect to get 20% of the cake on every interaction, regardless of whether he meets greedy or modest.

- Modest’s payoff: \((0.20 \times 0.25) + (0.20 \times 0.75) = 0.20\)
- Greedy’s payoff: \((0.80 \times 0.25) = 0.20\)

Modest gets 20% each time she encounters another modest, which is 25% of the time; and gets 20% each time she encounters a greedy, which is 75% of the time. Likewise, greedy gets nothing when he encounters another greedy, which is 75% of the time, but gets 80% each time he encounters modest, which is 20% of the time. Thus, both strategies can expect to receive, on average, 20% of the cake. This is a stable evolutionary equilibrium, since no other strategy can do better in this population.

The possibility of unfair polymorphisms poses a problem for Skyrms’s argument that the fair division strategy is most probable. What Skyrms must show is

that the fair division equilibrium is more likely than the unfair polymorphism. To do so, Skyrms relies primarily on the notion of positive correlation, which occurs when “each type tends to interact more with itself than would be expected with random pairing.” If positive correlation is likely, that will dramatically reduce the basin of attraction for the unfair polymorphism, and bolster Skyrms’s argument that the fair division strategy is most probable. For “greedy” will then encounter “greedy” more often and receive nothing, “modest” will encounter “modest” and only receive 20%, and “fair” will meet “fair” and receive 50%. Fair division will eventually spread. With this assumption Skyrms thus shows that fairness is most likely to emerge as equilibrium behaviour from replicator dynamics, and thus concludes that we have the beginning of an explanation for our fairness norms. According to him, the norms that we have are ones that have been selected through evolutionary processes.

**Cooperation and The Stag Hunt**

Let us now turn to Skyrms’s treatment of cooperation. In *The Stag Hunt and the Evolution of Social Structure*, Skyrms models cooperation by the Stag Hunt game. Rousseau introduces the story of the Stag Hunt in his *A Discourse on Inequality* as follows: “If it was a matter of hunting a deer, everyone well realized that he must

---

remain faithful to his post; but if a hare happened to pass within reach of one of them, we cannot doubt that he would have gone off in pursuit of it without scruple.”

Skyrms thinks the Stag Hunt models the relevant features of human cooperation better than the one-shot Prisoner’s Dilemma does. Skyrms argues further that the one-shot Prisoner’s Dilemma can be transformed into the Stag Hunt by repeating the game. Recall that in the one-shot Prisoner’s Dilemma there is only one non-cooperative equilibrium. When the game is repeated, different equilibria emerge and, thus, the “all defect” equilibrium is no longer the unique equilibrium of the game. When there is a possibility of being paired with the same individual again (or your past strategy selections become known to others), one’s reputation begins to matter. Consequently, the strategy of reaping the greatest rewards on a single shot (and thus defecting) begins to look less attractive, since the aim then becomes to secure multiple instances of cooperation.

In the two-person Stag Hunt, each player can either hunt stag or hunt hare. In order to hunt stag successfully, both players must choose that strategy. An individual

---


29 In particular, Skyrms thinks that the process of reforming the social contract is best represented as a Stag Hunt. In the case of reform, the state of nature is taken to be the existing social contract, which is an equilibrium. The new contract must also be an equilibrium if any reform is going to be successful. Skyrms thinks that this parallels the move from the hunt hare equilibrium to the hunt stag equilibrium; that is, from a riskless but less rewarding payoff to a riskier (presumably because any kinds of changes to the contract leave open the possibility of instability) and more rewarding payoff.

30 The Stag Hunt, 4. Repeating games like the Prisoner’s Dilemma changes the equilibrium structure of the game and makes available cooperative equilibria. I will leave the details of the implications of this for the next chapter. At this point all that is needed to proceed is to recognize that we are here transforming the game from a cooperation problem to an equilibrium selection (or coordination) problem.
may opt instead to go at it alone and hunt hare. Hunting stag yields a higher payoff than hunting hare. But the payoff of hunting hare is guaranteed, whereas the payoff associated with hunting stag is contingent on the other player’s cooperation. The Stag Hunt can be represented formally as follows.\footnote{Skyrms outlines the dynamics of this game in \textit{The Stag Hunt}, 1-13.}

<table>
<thead>
<tr>
<th>Player 1</th>
<th>Stag</th>
<th>Hare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stag</td>
<td>(4, 4)</td>
<td>(0, 3)</td>
</tr>
<tr>
<td>Hare</td>
<td>(3, 0)</td>
<td>(3, 3)</td>
</tr>
</tbody>
</table>

In this game, there are two equilibria: both hunt stag or both hunt hare. This lies in contrast to the single “defect” equilibrium in the one-shot Prisoner’s Dilemma, where, no matter what the other player does, the rational course of action is always to defect. While mutual cooperation in the Prisoner’s Dilemma yields a higher payoff than does mutual defection, this payoff is lower than what would be afforded to one who unilaterally defects. The Stag Hunt has a different structure. In the Stag Hunt, if player 1 suspects that player 2 will hunt stag, he will do the same. And if he suspects that player 2 will hunt hare, he will hunt hare. The set of strategies where both players hunt stag is “payoff dominant,” while hunting hare is the “risk dominant” strategy. The best outcome for each player is if each hunts stag. But this payoff is not guaranteed, and is contingent on the other player hunting stag. And hunting stag
leaves one open to the worst possible outcome, namely hunting stag while the other player hunts hare. Hunting hare, by contrast, will award a player a guaranteed payoff, but one that is lower than that achieved if both hunt stag. Thus, while in the Prisoner’s Dilemma the rational course of action is always to defect, in the Stag Hunt, the cooperative outcome is an equilibrium and is thus not strictly irrational (even if that equilibrium is riskier than the alternative).

When stag hunters meet hare hunters, the stag hunters receive a payoff of zero, while the hare hunters receive a payoff of 3. When two stag hunters meet, each receives a payoff of 4. When two hare hunters meet, each receives a payoff of 3 (not by any cooperative outcome, but merely because each obtains his or her own hare, which has a value of 3). Stag hunters will reap greater benefits when they pair with other stag hunters, but will only profit if they meet other stag hunters. Hare hunters, on the other hand, will profit regardless of whom they encounter, but will do so to a lesser degree than will stag hunters who meet other stag hunters. Given the fact that stag hunters will do worse when they randomly encounter hare hunters, and given that in such instances their payoffs will be less than those awarded to hare hunters, and that hare hunters will always receive a positive payoff, hare hunters will do on average better than stag hunters will, and will eventually come to dominate the population.

Skyrms says that “a pessimist, who always expects the worst, would hunt hare. But it is also true with these payoffs that a cautious player, who was so uncertain that he thought the other player was as likely to do one thing as another,
would also hunt hare. Hunting hare is said to be the risk-dominant equilibrium.”32 Thus, a puzzle arises when we seek to explain how a population can move away from a risk-dominant equilibrium.

The fact that human beings (and other creatures) do engage in cooperative endeavors33; that is, the fact that there are stag hunters in the world reveals a deficiency in the above descriptive account of the dynamics of strategic interactions. Skyrms says that “what is missing is an account of the evolution of the structure of interactions. Game theory takes the interaction structure as fixed. But in real life individuals adjust with whom they interact on the basis of past experience. This is a fundamental aspect of social behaviour that is completely absent from the theory of games.”34 What needs to be added to the account is a way to overcome risk-dominance in order to achieve the payoff-dominant outcome.

Skyrms appeals to the concepts of location, signaling, and association to explain how stag hunting may have evolved in a given population.35 These are mechanisms that generate positive correlation, and promise to change the dynamics of the game and push the equilibrium away from defection and towards cooperation. How readily other cooperators are able to interact with one another (and, thus, in the Stag Hunt, reap the rewards associated with both hunting stag) will depend on their


33 Technically speaking, hunting stag is not a “cooperative” outcome, since it imposes no cost on the actor and is an equilibrium of the game.


35 Skyrms treats these concepts at length in The Stag Hunt, 15-85.
location in relation to one another, their ability to communicate which strategy they plan to employ, and what kind of behavioural reinforcement mechanisms are employed. Stag hunting will only evolve if players have assurance that others will hunt stag. And this assurance can be acquired if we know that other stag hunters are close by and occupy a large portion of the population; if we can signal to others that we intend to hunt stag; and if cooperative behaviour can be reinforced (for example, if hunting stag is successful and others in a population adopt the policy of imitating the most successful strategy). With these things in place, Skyrms claims that we can see how stag hunting will soon come to spread throughout a population.

**The Need for Positive Correlation**

Positive correlation carries a heavy burden for Skyrms. Skyrms’s account of the evolution of fair division in the Divide the Cake game and cooperation in the Stag Hunt both rest on the plausibility that like strategies will pair with like strategies. But this assumption is questionable.

Skyrms’s argument for an evolutionary explanation of the evolution of fairness relies on being able to show that populations will tend to converge on the 50/50 division. More specifically, Skyrms must show that the 50/50 division is more likely than the unfair polymorphism. In order to do this, Skyrms adds the assumption of positive correlation—specifically, that like strategies will pair with like strategies. On this assumption, each strategy is more likely to interact with strategies like itself than would be predicted by random pairing. This subsequently enlarges the basin of
attraction for the 50/50 split, and carries the population away from the unfair polymorphism and towards fair division.

Martin Barrett et. al.\textsuperscript{36} note, however, that while the basin of attraction to get out of the polymorphism is enlarged with positive correlation, anti-correlation — where players pair with individuals who are unlike them (e.g., greedy pairs with fair, fair with modest, and so on) — enlarges the basin of attraction for the “unfair” polymorphism. Justin D’Arms further suggests that once we diverge from the assumption that interactions will be random, we open the possibility to other kinds of correlative encounters, including anti-correlation. As he says, “Skyrms’s correlation assumption amounts to a relaxation of the original rules of the model — individuals no longer pair off randomly. The more seriously we want to take the idea that the players in these games are rational agents, the more we should be tempted to look for mechanisms of learning from the environment other than mere imitation of strategies that have been successful in the past.”\textsuperscript{37} Thus, we may ask whether Skyrms is any more entitled to make his assumption about positive correlation than one would be to make an assumption about, say, anti-correlation.

Skyrms does say that positive correlation may be plausible, given the likelihood that individuals who enter into a population will be spatially located close


to one another and, thus, may interact more frequently with one another. This requirement is, however, hard to square with modern populations, which are highly diverse and mobile. But even if it were the case that we could grant that incoming members would remain in close proximity to one another, and thus that positive correlation could occur based on spatial location among members, it is not obvious that we can maintain the assumption of positive correlation when individuals are granted high-level cognitive abilities. Once we increase cognitive function we open the door to more strategic types of interaction. A “greedy” individual who is able to recognize other “modest” individuals and who is able to calculate the benefits of interacting more frequently with those individuals unlike himself may begin a trend of anti-correlative behaviour, which could plausibly be further propagated through imitation.

Nor is it clear that Skyrms’s appeal to correlation will explain why stag hunting will dominate the population. Skyrms relies on behavioural reinforcement to explain how stag hunting may spread throughout a population. Whether we get to a cooperative outcome will depend on how others learn to adopt different strategies. This mainly occurs through imitation of the best strategy: that is, individuals will adopt those strategies that yield higher payoffs than others in a given population. But

---


39 We will see the implications of modern human populations for the possibility of group selection in chapters three and four. Richerson and Boyd maintain that the prevalence of migration between populations dramatically reduces homogeneity at the genetic level. On this view, it is thus implausible to suppose that the correlation that Skyrms refers to above is primarily genetic in the human case.

the same mechanism that Skyrms uses here can be employed to explain how defection may arise: that is, if one individual defects and hunts hare and others imitate that, then they too will hunt hare, and this will in turn narrow the basin of attraction for stag hunting, and shift the dynamics back to hare hunting. In other words, while being suitably located with respect to other stag hunters and having appropriate signals to identify themselves to other stag hunters might be able to generate enough correlation to give to stag hunters the cooperative gains, it is plausible that in some cases these signals might fail (or stag hunters might be improperly located to one another) in such a way that they fail to meet and thus fail to reap the cooperative rewards. In such cases, the hare hunters would achieve a higher utility than stag hunters and, thus, if successful strategies are imitated, hare hunting would be replicated. In the absence of positive correlation, since hare hunters do better, it is not inconceivable that hare hunting would often be replicated. Skyrms’s account fails to explain convincingly why it is that we would not see more hare hunters.

This worry is exacerbated if we add in a bargaining problem, where having secured a stag, stag hunters have to divide the spoils. This makes stag hunting a particularly risky business. As Henry Ergas puts it:

In period one, agents get to choose whether to be hare hunters or stag hunters; in period two, they then encounter other agents, and if stag hunters encounter stag hunters, they get a combined pay-off; and then in period three, the stag hunters have to bargain about how to divide that combined pay-off. The effect of this is to make cooperation more risky, as the stag hunter now runs two
risks: the risk of meeting a hare hunter, and getting a pay-off of zero; and the risk that even if s/he meets a stag hunter, that stag hunter will play greedy in period three, reducing the agent’s return on cooperation to a lower share or possibly to zero.\(^{41}\)

This again makes cooperation appear fragile and, thus, if Skyrms’s account of how cooperation comes about is right, it would seem that we should not expect to find many examples of stag hunters.

Even if Skyrms’s account can explain the transition from hare hunting to stag hunting, it is not able to explain other forms of social cooperation. Skyrms claims that the Stag Hunt is the best model of human cooperation. But the Stag Hunt model suffers from a significant deficiency that precludes it from such a role. As Matthew Simpson has argued, Skyrms’s use of the Stag Hunt to model human cooperation fails to accommodate the problem of free riding—that is, of reaping the rewards of cooperation without bearing any of the costs. Simpson articulates how this is a problem for Skyrms:

The difficulty is rooted in the structure of the game itself. The assumption behind the stag hunt is that it can succeed only when there are zero defectors. But in most cooperative enterprises, this is not the case. Success does not rely on everyone cooperating. This is the fact that opens the door to the free rider, for whom the model of the stag hunt has no room….A model of the problem

---

of cooperation should include the problem of the free rider within the stag hunt.42

**Generalizability of EGT Explanations**

However successful Skyrms may be in explaining the shift from hare hunting to stag hunting in the Stag Hunt game (or fair division in the Divide the Cake game) in terms of EGT, it does not follow that we can extrapolate from these models to human society. As was mentioned in our discussion of the Divide the Cake model, players with high-level cognitive abilities may behave differently from how players behave when they have no or low-level cognitive abilities attributed to them, as they do for the purposes of EGT. Relevant differences can be expected as soon as individuals are in situations in which they realize there are advantages to be gained from anti-correlation, and understand that they can bring that about by deception and free riding. As reason is increased, anti-correlation can also be expected to increase. Thus, as players become more like human beings (with respect to cognitive function), the less we can rely on correlation to generate the behaviour Skyrms predicts. In the absence of a foolproof device to de-incentivize defection, cooperation among selfish utility-maximizers cannot be guaranteed.

Skyrms clearly understands that his models are simplifications, but he also thinks that they can be extended. As he puts it:

---

The fundamental techniques and principles [of correlated interactions] surveyed may plausibly be applied to other species as well as man, and at various levels of biological organization. The considerations of location, signaling, and network formation, introduced here in their simplest forms, are capable of being combined to form models of complex phenomena.\textsuperscript{43}

The problem is to understand how Skyrms can extend his models. For, as Herbert Gintis says, “By assuming that agents have very low-level cognitive capacities, evolutionary game theory ignores one of the most important of human capacities, that of being able to reason. Human society is an evolved system, but human reason is one of the key evolutionary forces involved.”\textsuperscript{44} It is thus hard to see how Skyrms can account for human society without giving up his project of explaining it in terms of EGT.

In the next chapter we will turn to Binmore’s account. He, like Skyrms, contends that evolutionary processes can help to solve equilibrium selection problems. But Binmore does not remove reason from his analysis of cooperation problems and employs RGT to solve them. I will argue, however, that this approach too is unsuccessful, and that Binmore is unable to provide an adequate explanation of the emergence of cooperation among self-regarding rational agents.

\textsuperscript{43} Skyrms, \textit{The Stag Hunt}, xiii.

CHAPTER TWO

The Folk Theorem and Cooperation in Repeated Games

In the last chapter we saw that RGT’s limitations in predicting or prescribing behaviour in both coordination and cooperation problems have led some to take up an EGT analysis of those games. We examined Skyrms’s account of how EGT might be employed to assist with problems of coordination and cooperation. I argued that Skyrms’s reliance on EGT as his chief descriptive tool is unable to capture crucial aspects of human cooperation. More specifically, he fails to account for the implications that high-level cognitive functions of human beings will have on the possibility of cooperation. The introduction of higher levels of reason introduces the possibility of free rider strategies and, thus, destabilizes cooperation.

In this chapter I will turn to Binmore’s project of naturalizing the social contract. Like Skyrms, Binmore turns to evolution to resolve coordination problems. But Binmore handles the cooperation problem differently. He maintains the *homo economicus* model of man, according to which individuals are assumed to be selfish utility-maximizers, and argues that social interaction ought to be modeled as a repeated game. Binmore employs the folk theorem of repeated games to argue that cooperative equilibria are available in repeated games and, thus, he contends that it is possible to reconcile cooperation with the conception of individuals as rational utility-
maximizers. In what follows I will examine his account, and will argue that the folk theorem is not a sufficient mechanism to generate and maintain large-scale cooperation.

**Equilibrium Selection and The Game of Life**

The first step Binmore takes towards solving the equilibrium selection problem is by modeling the game that individuals in a society are engaged in. He refers to this game as the “Game of Life.”¹ Naturalists like Binmore maintain that our evolutionary history is important; that is, they think our current social contract (or at least components of it) likely originated from the interactions of our early human ancestors and, thus, an adequate explanation of the social contract must be sensitive to that history. Binmore claims that the Game of Life is an indefinitely repeated game.² This is so because he speculates that the interactions between our early ancestors from whom we inherited our cooperative norms were likely members of small hunter-gatherer groups and, thus, individuals interacted in repeated rather than one-shot games. The players in the game are the individual members of society, and

---

¹ In Binmore’s words:

The Game of Life may change over time, but it is beyond our power to alter its rules at will. Its rules are determined by the laws of physics and biology; by geographical and demographic facts; by technological and physiological constraints; and by whatever else sets unbreakable bounds on our freedom of action. *Game Theory and the Social Contract, Volume 1: Just Playing* (Cambridge, MA: MIT Press, 1994), 5.

² Binmore says of Gauthier’s solution to the problem of cooperation modeled by the one-shot Prisoner’s Dilemma (to which we will turn in chapter five) that he has given “a wrong analysis of the wrong game.” Binmore says that if social interaction did conform to the one-shot Prisoner’s Dilemma model, cooperation would have never evolved. Ken Binmore, *Game Theory and The Social Contract, Volume II: Playing Fair* (Cambridge, MA: MIT Press, 1998), 174.
are presumed to be selfishly interested. The sets of strategies available for individuals are those that constitute equilibria to the Game of Life.

There are a number of different equilibria in the Game of Life. These equilibria are different ways in which interaction can take place in various different situations or “games.” Some of them might be more mundane than others—such as (to use Binmore’s examples) who is to wash the dishes tonight, or how to divide a scarce dish at a dinner party. Others are more significant—such as how to allocate resources or efficiently work together to generate a surplus. We can decide how to do these activities in a number of different ways. Each society has a set of conventions to determine this. These conventions are often equated with the rules of morality (or at least a subset of these conventions is identified as moral). Binmore thinks that these rules evolved to help us to settle on one particular equilibrium over another. It doesn’t really matter if the one who cooked the dinner washes the dishes or the one who was served the dinner does; all that really matters is that someone does so. And having rules to quickly decide is better than not having such rules. He suggests that the need for this probably started in the family group with smaller matters and then expanded to the larger social sphere. A society’s social contract is thus, “the set of all the commonly understood coordinating conventions operated by a society.” Binmore thinks that cooperative (or fairness) norms are one such set of rules for coordinating on an equilibrium.

---


5 Binmore, *Just Playing*, 5.
Binmore distinguishes the Game of Life from what he calls the “Game of Morals.” The Game of Life refers to the game that is actually being played by individuals. The Game of Morals, by contrast, is an equilibrium selection device that can be used by players to coordinate on a particular equilibrium in the Game of Life. The Game of Morals is “an idealized representative of a class of equilibrium selection criteria washed up on the beach along with the human race by forces of biological and social evolution.” According to Binmore, we already use something to coordinate our behaviour, and the Game of Morals is an idealization of these methods. The Game of Morals is an imaginary game that Binmore thinks that we play when determining what social contract should be brought about or solving problems of distribution where there is no clear convention. Binmore contends that the mechanism we use to determine fair outcomes is akin to the device used in Rawls’s original position.

This device asks us to imagine ourselves behind a “veil of ignorance” that conceals our identity from ourselves. We then make contracts supposing that we have an equal chance of being any of the players. But while Rawls appeals to reason to justify appeal to the original position, Binmore looks to evolution to explain why we already use this device. He says,

---

6 Binmore says that, “to carry out their part in sustaining a social contract, the citizens of a society need rules to regulate their conduct....I take the totality of all such rules as defining the morality game for a particular society.” Just Playing, 5-6.

7 Binmore, Playing Fair, 9.
I think that the reason most people find the device of the original position intuitively attractive as a fairness criterion has nothing to do with the Kantian arguments offered by Harsanyi and Rawls. I believe that its appeal lies in the fact that we recognize it as a stylized version of a principle that we already unconsciously apply every day when interacting with our peers. From such a perspective, fairness is interpreted entirely in naturalistic terms. The original position is merely a device that has been washed up on the beach along with the human race by the forces of biological and social evolution.\(^8\)

Once we understand this device that we use to adjudicate such coordination problems and select equilibria from the number of feasible outcomes in everyday life in the Game of Life, Binmore thinks that we can apply it to the larger sphere to more important social matters, such as moral problems and reforming the existing social contract. In Binmore’s view, the possibility of reforming the social contract (that is, of moving from one equilibrium to a Pareto-superior equilibrium, i.e., an equilibrium where at least one player is made better off without worsening the position of any other) depends on whether such a reform is feasible for the parties concerned. Any kind of recommended reform that is not in keeping with the way that humans actually behave in their cooperative endeavors is not likely to succeed. We must, therefore, take human beings as they are, and narrow the scope of recommended reforms to those that are most likely to succeed.\(^9\) For Binmore, “the defense for such a proposal is entirely pragmatic. Here is a tool supplied by Nature. Let us use it to improve our

---


lives, just as we use whatever tools we find in our toolbox when making repairs around the house.”

The success of Binmore’s solution to the equilibrium selection problem, the content of the norms he claims to naturalize, and how to invoke reforms lies beyond the scope of my aims in this thesis. Whether or not he succeeds in these endeavors ultimately is secondary to whether he can explain compliance. In fact, nothing can be done with regard to reform unless the proposed mechanisms supporting cooperation are accurately captured. As we will see, Binmore’s argument relies on the folk theorem, according to which repeated games leave the possibility open for cooperative equilibria. According to Binmore, we have an evolved fairness norm to help to select an equilibrium. But if the conditions for the folk theorem do not obtain, then the cooperative equilibria will not emerge. Thus any solution given to the equilibrium selection problem will depend on which equilibria are available. And if Binmore is wrong about the folk theorem, then his solution to the equilibrium selection problem, along with any recommendation for reform, will not have practical effect. In the remainder of this chapter I will argue that the mechanisms Binmore proposes are not sufficient to explain widespread cooperation.

The Folk Theorem and Compliance

According to Binmore, evolution thus gives us the tools for equilibrium selection and, consequently, the principles that govern society. But self-interest

10 Binmore, “Justice as a Natural Phenomenon,” 3.
ensures compliance to these rules. Binmore relies on the folk theorem\textsuperscript{11} of repeated games to explain this. The social contract is a set of conventions that are used to coordinate on equilibria. Different conventions are employed for different selection procedures. Moral conventions are a subset of those employed to arrive at “moral” or cooperative equilibria. These cooperative equilibria become available when we introduce repeated games.

We have seen that it is not always the case that the equilibrium of a particular game is the mutually advantageous equilibrium. We have seen, for example, that the one-shot Prisoner’s Dilemma has only a single, non-cooperative equilibrium, where each party ends up doing worse by playing the equilibrium strategy than they would have had they both cooperated. Repeated games, however, change the equilibrium structure of many one-shot stage games. When certain games are repeated, cooperative equilibria emerge in games that weren’t present in the one-shot version. As Gintis puts it: “When a game $G$ is repeated an indefinite number of times by the same players, many anomalies associated with finitely repeated games disappear. Nash equilibria of the repeated game arise that are not Nash equilibria of $G$.\textsuperscript{12} The folk theorem of repeated games provides us with an analysis of these equilibria. Binmore says that “for indefinitely repeated games, the folk theorem tells us that we


\textsuperscript{12} Gintis, \textit{Game Theory Evolving}, 201.
do not need to rely on anything but the enlightened self-interest of sufficiently forward-looking players to maintain the full panoply of cooperative possibilities.”

Let us look more closely at how the folk theorem works by returning to the Prisoner’s Dilemma. The one-shot version of the Prisoner’s Dilemma has only one equilibrium: the non-cooperative outcome where both players defect. In the Prisoner’s Dilemma, the cooperative outcome is not an equilibrium and, thus, cooperation runs up against self-interest (and rationality). We saw in the last chapter that repeating the game changes the equilibrium structure. We call each period of the game in the repeated game a “stage game.” Each player’s payoff for the repeated game becomes a sum of the payoffs received in each instance of the stage game, weighted by a discount factor. Rational players then base their strategies in each stage game on the past moves of their opponents. Thus, when the Prisoner’s Dilemma is repeated, one’s reputation begins to matter. In order to secure the best possible outcome for oneself, one will base one’s strategy on what he thinks the other player will do. The other player will do likewise. When games are repeated, the possibility of retributive punishment also emerges. This can take the form of either a withdrawal of future cooperation or more targeted and possibly extreme punitive measures.

Binmore describes how repeating the game changes the situation as follows:

Consider, for example, the indefinitely repeated Prisoners’ Dilemma. This is very different from the one-shot version of the Prisoners’ Dilemma studied so far. In the repeated version, Adam and Even play the Prisoners’ Dilemma over

---

13 Binmore, Just Playing, 293.
and over again until some random event intervenes to bring their relationship to an end. It is usual to model this random event by postulating that each time they finish playing a round of the Prisoners’ Dilemma, there is a fixed probability $p$ that they will never play again. Interest then centers on the case when $p$ is very small, so that the players will have good reason to believe that they have a long-term relationship to nourish and preserve.\footnote{Binmore, \textit{Playing Fair}, 115.}

If I know that if I defect then my partner will subsequently defect on me, I must factor that into my utility calculation. While a single instance of unilateral defection might yield a greater payoff than will cooperating (as it does in the case of the one-shot Prisoner’s Dilemma), a single instance of defection followed by the punishment that I will incur from my partner (usually consisting a series of defections) will yield a lower payoff than would be afforded by a series of acts of mutual cooperation. Thus, “the strategic considerations in an indefinitely repeated game are totally different from those in a one-shot game because the introduction of time permits the players to reward and punish their opponents for their behavior in the past.”\footnote{Binmore, \textit{Playing Fair}, 115.} According to Binmore, “the problem for evolution in creating a cooperative species is therefore not that there are no cooperative social contracts available as Nash equilibria in our indefinitely repeated game of life, but that there are an embarrassingly large number of such equilibria.”\footnote{Binmore, “Justice as a Natural Phenomenon,” 7.}
The Robustness of the Folk Theorem

The folk theorem thus states that repeated games introduce the possibility of players landing on a cooperative equilibrium in a game that, when only played once, lacks such a cooperative equilibrium. Individuals will cooperate with those who have cooperated in the past, and will punish those who have defected. Gintis outlines the requirements of the folk theorem as follows:

The folk theorem asserts that if the signals of defection (that is, the signals that a player deviated from the behavior specified by the Nash equilibrium) are sufficiently high quality, and if players have sufficiently long time horizons, the repeated game based on [a game] $G$ can attain Pareto-efficiency, or at least approximate Pareto-efficiency as closely as desired.\textsuperscript{17}

Thus, cooperative outcomes can be sustained in repeated games among rational individuals, so long as individuals have an adequately low discount factor (that is, they do not discount the future so much that future payoffs do not factor significantly into their utility calculations) and, most importantly, so long as there is adequate public information about who cooperated. When these conditions are not met, cooperation becomes unstable. Let us now turn to examine the likelihood of these conditions obtaining.

\textsuperscript{17} Gintis, Game Theory Evolving, 201. Binmore defines “Pareto Efficiency” to mean “nothing gets wasted,” and goes on to say that “economists follow Vilfredo Pareto in taking the absence of waste to be equivalent to the requirement that nobody can be made better off without someone else being made worse off.” Natural Justice, 7.
Discounting the Future

We can begin with the condition that individuals must be sufficiently forward-looking, i.e., must have sufficiently low discount factors. Without this condition, individuals will discount the payoffs associated with future cooperation in such a way that the threat of punishment for defection will prove ineffective. If the prospect of punishment through withdrawal of cooperation is what sustains cooperation, then cooperation will only be maintained if individuals care enough about their future payoffs.

But as Gintis points out, it is most likely that our early ancestors cared less about the future than would be required to sustain cooperation through the threat of retaliation.\(^\text{18}\) Life for our ancestors was likely risky, which imposes an uncertainty about one’s future. And if your future life is uncertain, you will likely be consuming goods while you have the chance, rather than forgoing immediate gains and instead planning far ahead into the future. Thus, it seems that the requirement of forward-lookingness is at odds with the conditions under which our early ancestors likely lived, and from which cooperation is derived.

Public Information and Punishment Structure

The most essential requirement for the folk theorem is that information be public. That is, to guard against defectors, information about who is a cooperator and who is a defector needs to be both readily accessible and accurate. However, this

requirement becomes more difficult to meet as group size increases. As Gintis notes, “in most situations in real life, the larger the number of players participating in a cooperative endeavor, the lower the average quality of the cooperation-vs.-defection signal because generally a player observes only a small number of other players with a high degree of accuracy, however large the group involved.”19 This (I will now argue) undermines the plausibility of appealing to the folk theorem to explain cooperation in human society.

The folk theorem relies on the efficacy of punishment. There are two ways in which an individual who shirks can be punished by the group. The first is through a withdrawal of cooperation. The second is through active and targeted punishment. Gintis says of the first method of punishment that “the folk theorem models that have plausible stability properties are those in which, when shirking is detected, the group reverts to noncooperation for a sufficiently large number of periods that it is not profitable to shirk. These are called ‘trigger strategies.’”20 In order to punish in this way, however, a sufficient number of individuals needs to get on board with withholding cooperation with the shirker. And in order for that to be successful individuals will need to communicate with one another to identify the shirker. That seems like something that could readily be done in smallish groups. But as group size increases it seems that the chances that one could shirk and not subsequently be reliably identified and thus punished diminish. Furthermore, when individuals

---


withdraw punishment as they do on this model, everyone forgoes the benefits of cooperation. Thus since this kind of punishment is ultimately costly to everyone, this increases the chances that shirkers will go unpunished. As Richerson and Boyd point out, “if reciprocators use the rule, only cooperate if all others cooperate, [a] defector induces other reciprocators to stop cooperating. These defections induce still more defections. Innocent cooperators suffer as much as guilty defectors when the only recourse to defection is to stop cooperating. On the other hand, if reciprocators tolerate defectors, then defectors can benefit in the long run.”

Gintis says further that:

In small groups with highly accurate public signals, trigger strategy models are quite robust. However, in larger groups with private or imperfect signals, these models lead to very low levels of cooperation. It is not surprising, then, that trigger strategies are rarely observed to be the strategic mechanism through which cooperation is maintained in empirical studies of hunter-gatherer societies or other forms of social cooperation involving more than a few agents. The inefficiency of trigger strategies is due to the fact that all group members are punished for the sins of each.

It thus seems implausible that large-scale cooperation among self-regarding agents can be explained by the folk theorem and a punishment structure consisting in the withdrawal of cooperation.

---


Targeted punishment promises to resolve the above problem about the inefficiency of withdrawing cooperation.\textsuperscript{23} A single individual (or group of individuals) might, for example, take on the role of punishing a shirker. In this way, information about the shirker need not be as public and reliable as it does in the first case. Nor does the entire group suffer by reverting to a non-cooperative strategy. And if the punishment is sufficiently severe (for example, one might get kicked out of the group for shirking), that will provide an incentive for others to cooperate.

But this raises what Gintis refers to as a “second-order free rider problem,”\textsuperscript{24} for we must now ask why any self-regarding agent would punish a shirker. Punishment is effective in holding the folk theorem together only if it will be rational to punish, and that will be rational only if punishers can benefit from that punishment. Not only is it not clear that there is a benefit associated with being a punisher, punishment is in itself costly. There is no reason to think that self-interested people will want to take on the role of punishers, and punishing those who do not punish threatens to lead to an infinite regress.\textsuperscript{25} And since a potential punisher has no (self-interested) reason to be a second-order punisher unless she has reason to think she will be punished if she does not, and since everyone is in that same position, no self-interested person will have a reason to be a second-order punisher. Nor do we have to rely on an \textit{a priori} argument to be skeptical about the efficacy of second-order punishment. Gintis reports that, “while there is a high frequency of punishment of


\textsuperscript{24} Gintis, “Behavioral Ethics Meets Natural Justice,” 11-12.

\textsuperscript{25} For who will then punish those who fail to punish those who fail to punish? And so on.
norm violators in social groups, second-order punishment is virtually never observed. An individual who refuses to participate in punishing a malefactor is simply left in peace.”

Gintis says further:

Real-world social relationships must have evolved historically under adverse and primitive conditions, and must be capable of withstanding invasions by mutant strategies. By contrast, repeated game models with many agents are…dynamically unstable and tend to fail when signals are noisy and private, and it has never been shown that they could be repaired to have the dynamic stability properties that render the evolution of cooperative institutions possible and ensure their structural continuity through time.

In other words, what made cooperation possible must have been stable. Repeated games, as I have argued, are not. Therefore, repeated games do not lie at the foundation of cooperative behaviour.

Experimental Game Theory and Homo Economicus

Binmore’s account of cooperation also faces a problem from experimental game theory. Gintis sees a lot of discontinuity between experimental evidence and the mechanisms that Binmore proposes to explain cooperation. According to the folk theorem, cooperation is held together by self-interested considerations of sufficiently forward-looking rational individuals. It should follow from this that we would see a

---


lower level of cooperation in one-shot interactions, since individuals need not be concerned with the possibility of retributive punishment for defection. But in experimental settings, the level of cooperative behaviour far exceeds what self-interested assumptions about human beings would predict. For example, experimental results show high levels of cooperation in cases where it would be predicted that self-regarding agents would not cooperate, such as in one-shot anonymous versions of the Prisoner’s Dilemma. Gintis reports:

Laboratory experiments reveal forms of prosocial behaviour (e.g., rejecting “unfair” offers in an ultimatum game, or punishing free riders in a public goods game) that relate directly to questions of justice and fairness, yet contradict the *Homo Economicus* model. The notion that human sociality can be explained by “enlightened self-interest,” even when accompanied by respect for the Original Position, will not likely survive a close study of the evidence.²⁸

Binmore attributes the higher than predicted levels of cooperation in these cases to human error—that is, to individuals’ failure to recognize that they are indeed playing one-shot anonymous games. To explain the observation that more than 50% of subjects cooperate in one-shot anonymous Prisoner’s Dilemmas, Binmore says:

We think that the most likely explanation is that the framing of the game triggers a social norm that the players are accustomed to using when going

---

about their everyday affairs. We see such a social norm as an equilibrium selection device that has evolved to allow us to coordinate our behavior in one or more of the games we play in real life. To survive, it must select a Nash equilibrium of the real-life games for which it is adapted, but the strategy profile it selects need not be a Nash equilibrium of a game the subjects face in the laboratory. In the case of laboratory games such as the prisoner’s dilemma that model the private provision of public goods, the relevant real-life game is an indefinitely repeated game, for which the folk theorem tells us that cooperation can be sustained as a Nash equilibrium by strategies that punish anyone who defects.29

Thus Binmore explains cooperation in Prisoner’s Dilemmas in experimental situations in terms of a mistaken application of a strategy adapted for a different game in the Game of Life. And he goes on to contend that subjects will learn to no longer cooperate once the game is repeated a few more times.30

However, as Gintis reports, even when individuals are explained the rules of the game and asked to repeat the rules to ensure that they understand them, cooperative levels are much higher than they ought to be if individuals were indeed selfishly motivated. Gintis says that “if the self-regarding models were correct, it is not clear why humans would ever cooperate or punish in one-shot, anonymous interactions, where the carefully constructed incentives for self-regarding cooperation are conspicuously absent.”31

30 Binmore, “Why do People Cooperate?” 86.
A Better Explanation?

We have seen that cooperation in repeated games among self-regarding agents depends on punishment of defectors. Punishment of defectors requires that information about defectors be public and accurate, and that the punishment structures themselves be stable and efficient. In the case of punishment where all group members withdraw from future cooperative interactions with the shirker, information must be public and very reliable. Anyone who has ever played the telephone game will know that group size has tremendous effects on the reliability of information. Punishment in this sense might be effective in small groups where the informational demands can more easily be met, but as group size increases, the efficacy of punishment decreases and so too does the stability of cooperation. Targeted punishment has fewer informational demands but has trouble explaining why self-regarding agents would take on the role of punishers. I thus conclude that Binmore’s appeal to the folk theorem fails to account for the cooperation we find in society.

If cooperation requires a punishment structure that goes beyond withholding cooperation from defectors, and second-order punishment is not plausible, Gintis concludes that, “in the absence of second-order punishment, punishers must be altruistic in one form or another.”32 He says:

The technical problems involved in developing an equilibrium with a high level of cooperation and assuming private information and self-regarding

---

agents are extreme. If punishing an observed breech of cooperative norms is costly, and if team members generally do not know which members observed which breeches, costly first-order punishment will not occur because those who see the defection know that they will not be punished for failing to punish. Therefore, first-order punishment must fail to be costly. There are various ways of achieving this result involving the use of mixed strategy sequential equilibria, so these models are vulnerable to the critique that the mechanisms involved are not seen empirically and they have very poor stability properties.33

As an alternative explanation of cooperation Gintis proposes the hypothesis of “strong reciprocity,” which he defines as “a predisposition to cooperate with others, and to punish those who violate the norms of cooperation, at personal cost, even when it is implausible to expect that these costs will be repaid."34 If individuals have the prosocial dispositions they are assumed to have on the “strong reciprocity” hypothesis, which includes a propensity to cooperate as well as to punish shirkers, then even in the absence of reliable information, cooperation can be stabilized. A disposition to cooperate will reduce the number of transgressions. And this, coupled with the propensity of others to punish transgressors, will alleviate the problems of motivating second-order punishment that are present among entirely self-regarding agents, and result in the punishment and subsequent deterrence of the (fewer) transgressions that will happen. According to Gintis, “a small amount of other-

regarding behavior (in particular, the willingness to punish defectors) dramatically improves the efficiency and stability of models of cooperation, rendering them fully capable of explaining cooperation in large groups even under adverse informational conditions.”

Strong reciprocity thus promises to provide an explanation of the emergence of cooperation by appealing to other-regarding preferences to ensure stability. This hypothesis is one that is easy to square with experimental evidence and gives a plausible explanation of the emergence of high levels of cooperation in large groups that is able to withstand the pressures on informational accuracy that an increase in group size produces. But whether strong reciprocity ultimately is a plausible explanatory mechanism depends also on whether it is compatible with an evolutionary explanation. Strong reciprocity requires that individuals do things for the sake of others at a cost to themselves, i.e., be altruistic. But altruism appears to be at odds with natural selection. In the next two chapters we will turn to examine whether it is plausible to think that evolution could have produced the tendencies required by strong reciprocity.

---

Thus far we have seen that Skyrms and Binmore both explain the emergence of cooperation in terms of the individual advantage that cooperative behaviour confers. Cooperators do better than non-cooperators so long as there is a reliable way to distinguish shirkers from cooperators. I argued that neither Skyrms nor Binmore provides us with a plausible account of such a mechanism and, thus, theirs fall short of an explanation of the emergence of cooperation in humans. Furthermore, what Skyrms and Binmore aim to provide, essentially, is an explanation of why it is in one’s interest to be cooperative. This makes the cooperation that we see only apparent. However, individuals seem to perform actions that are genuinely fitness-decreasing, which invites the question whether this behaviour can be rendered compatible with the standard tenets of evolutionary processes.

In the last chapter we saw that one way to account for experimental and observed behaviour is to appeal to strong reciprocity. Individuals disposed to strong reciprocity have a willingness to punish defectors. Since punishment imposes a cost on the actor and benefits others, a disposition to punish requires that individuals display a certain degree of altruism—that is, that they be disposed to act in ways that are contrary to their own interests. This chapter will examine whether nature could
produce this kind of disposition. Sober and Wilson think so. They contend that altruism—*genuine* altruism, i.e., behaviour done for others that lowers the individual’s fitness without genetic or other recouping—does exist and can be explained in terms of group selection. In this chapter I will examine their account. I will argue that group selection promises to be a useful framework from which to approach the explanation of the emergence of cooperation among humans, but it is not clear that Sober and Wilson’s emphasis on *biological* group selection serves as a sufficient explanatory mechanism.

**Varieties of Altruism**

Altruism is commonly understood to impose a cost on the individual who performs it and confer a potential benefit to the recipient. There is an important distinction to be drawn between evolutionary and psychological altruism.¹ Evolutionary altruism refers to an act that reduces the reproductive fitness of an individual, and increases that of another. This sacrifice needn’t be (and often is not) conscious, and is displayed by humans as well as mindless creatures. Psychological altruism, by contrast, concerns motives, and requires that an agent act contrary to his or her desires with the intention of providing a benefit to another individual in a non-instrumental way. Sober and Wilson put this distinction as follows:

The concept of psychological altruism is, in a sense, the mirror image of the evolutionary concept. Evolutionary altruism describes the fitness effects of a behavior, not the thoughts or feelings, if any, that prompt individuals to produce those behaviours. In contrast, psychological altruism concerns the motives that cause a behavior, not its actual effects. If your treatment of others is prompted by your having an ultimate, noninstrumental concern for their welfare, this says nothing as to whether your actions will in fact be beneficial. Similarly, if you act only to benefit yourself, it is a further question what effect your actions will have on others. Psychological egoists who help because this makes them feel good may make the world a better place. And psychological altruists who are misguided, or whose efforts miscarry, can make the world worse.²

A particular act can be evolutionarily altruistic in the sense that it lowers an individual’s fitness, and at the same time can be psychologically egoistic in the sense that it is selfishly motivated. Likewise, psychologically altruistic acts may work to one’s fitness advantage. My concern in this chapter will be primarily with evolutionary altruism. The existence of evolutionary altruism poses the same kind of explanatory difficulty for the evolutionary biologist as cooperation in Prisoner’s Dilemma situations does for the rational game theorist. Recall that cooperation in the

² Elliott Sober and David Sloan Wilson, “Summary of: Unto Others: The Evolution and Psychology of Unselfish Behavior,” Journal of Consciousness Studies 7, no.1-2 (2000): 185-6. Reprinted in Evolutionary Origins of Morality: Cross-Disciplinary Perspectives, ed. Leonard D. Katz (Thorverton: Imprint Academic, 2002), same pagination. Whether psychological altruists actually do exist lies beyond the scope of this thesis. However, it seems that if they do exist, they are very rare. There is a host of examples where it appears that one is acting in a psychologically altruistic way, but where one’s behaviour is actually self-serving. For e.g.: Unilever operates a tea plantation in Africa. They provide to their workers high quality health care and do much by way of reforestation efforts in East Africa. This looks altruistic. But it is costly and troublesome to train workers, so Unilever puts efforts into keeping the ones it does have healthy, and deforestation affects rain patterns, which affects tea production. Their “altruistic” behaviors actually turn out to be part of a well-calculated business plan.
one-shot Prisoner’s Dilemma conflicts with what rational utility-maximization would predict. It thus seems that, insofar as we are rational, cooperation in such situations would not occur. Likewise for the existence of altruism in nature. Since evolutionary altruism requires that an individual sacrifice his or her reproductive fitness in order to enhance that of another individual, and since natural selection works against fitness-decreasing characteristics, it seems that evolutionary altruism ought to have been eliminated through evolutionary processes. But altruism in nature is apparent, which invites the question how such behaviour could have emerged. As David Sloan Wilson puts it,

This is the paradox that makes altruism such a fascinating subject for evolutionary biologists. As humans we would like to think that altruism can evolve, as biologists we see animal behaviors that appear altruistic in nature, yet almost by definition it appears that natural selection will act against them. This is the sense in which evolution appears to be an inherently selfish theory.³

### Individual- and Group-Level Explanations

Individual-level explanations of the emergence of traits appeal to the fitness advantages that those traits confer onto the individual.⁴ A trait that evolves through

---


⁴ More precisely, confers a benefit to the individual, broadly construed. I say “broadly construed” here because it is not the benefits exclusively to the individual that count here, but the benefits conferred onto the genes possessed by the individual. In other words, natural selection operates at the level of
Individual selection has a fitness advantage within the group to which it belongs. Individuals who possess a trait that provides them with a fitness advantage over other individuals in that same group will outcompete those individuals. The fitness-enhancing trait will subsequently become more prevalent within the population.

Altruism imposes a fitness cost to the actor and confers a benefit to the recipient. But on that definition, altruists will necessarily move towards extinction within a group. According to Sober and Wilson, individual-level explanations of the emergence of altruism require that the cost of altruism be recouped elsewhere and, thus, fail to provide an explanation of genuine altruism. That is, parental care and helping one’s relatives are really instances of selfish behaviour: these are instances of behaviour that benefits the individuals themselves or their genes.

We might be tempted to conclude from the above that all cases of apparent altruism are only apparent. Once properly understood, the behaviour in question is actually individually advantageous, broadly construed, where the behaviour is beneficial either to the individual organism or the gene responsible for the behaviour. This, Sober and Wilson say, is what made the hypothesis of group selection appear superfluous, explaining that: “You don’t need the hypothesis to explain what you observe. Altruism is only an appearance. Dovishness isn’t present because it helps the group; the trait is maintained in the population because individual doves gain an advantage from not fighting to the death.”

As Randolf Nesse says,

---

5 Sober and Wilson, “Summary of: Unto Others,” 188.
From this inexorable logic, many have reached the bleak conclusion that altruism is impossible. Our actions, this argument goes, are products of a brain that is designed to advance the interests of our genes. Taking care of our children, and helping people who help us, may seem altruistic, but such actions actually help our genes. Other instances of apparent altruism can be attributed to manipulation, coercion, novel environments, mistakes, or conscious decisions to oppose the dictates of natural selection.6

Sober and Wilson contend that altruism—*genuine* altruism, i.e., behaviour done for others that lowers the individual’s fitness without genetic or other recouping—does exist and can be explained in terms of group selection. Where individual selection explains the emergence of traits in terms of selection occurring *within* the group, group selection explains them in terms of selection occurring *between* groups. According to the group selection hypothesis, some behaviours or traits evolved, not because they were advantageous to particular individuals, but because members of groups containing those traits did better than members of groups that did not.

Thus, while individual selection will favor the evolution of selfishness within groups (and thus altruists will be less fit relative to non-altruists within a single group), matters are different at the level of groups. Altruists will do, on average, worse than selfish individuals within the same group. But, so the argument goes, members of groups of altruists will do better than will members of selfish groups.

---

And if we grant that selection can occur at the level of groups, then we can explain the existence of altruism in terms of it: altruism evolved because members of altruistic groups did better than members of non-altruistic groups. Group selection thus promises an explanation of the emergence of behaviours or traits that genuinely reduce the fitness of an individual within a particular group, so long as those behaviours or traits increase the fitness of the members of the group that contains them relative to members of other groups which do not.

But how plausible is this? Let us say that for every recipient of an altruistic act, an individual gains 2 Darwinian fitness points. Every altruist loses 1 fitness point for every act of altruism she performs. When two altruists meet, each gains 2 but loses 1. When an altruist meets a selfish individual, she loses 1 and confers a benefit of 2 onto the selfish individual. When two selfish individuals meet, neither gains and each receives a payoff of 0.

Of two populations, one containing all altruists, the other containing all selfish individuals, the population of altruists will be fitter than the selfish group, since altruists each receive 1 point when they encounter one another, and selfish individuals receive nothing. If we translate these fitness points to number of offspring, we can see that the proportion of altruists in the global population (viz., the population resulting from the combination of the two groups) will increase.

However, in groups containing both altruists and selfish individuals, even if altruists make members of the group that contains them more fit than members of the group that does not, within the group, altruists will be less fit than non-altruists. A
selfish individual will do better against an altruist within a particular group. Selfish individuals will have higher levels of relative fitness than will altruists of the same group, and will thus have more offspring than altruists. Selfish individuals will prosper, and altruists will begin to go extinct. Thus, it appears that so long as selfish individuals are present in a population containing altruists, altruists will be at a selective disadvantage and will tend towards extinction.

**Simpson’s Paradox**

Sober and Wilson appeal to Simpson’s Paradox to explain how the evolution of altruism in mixed populations like the above is possible. Simpson’s Paradox “refers to the phenomenon whereby an event $C$ increases the probability of $E$ in a given population $p$ and, at the same time, decreases the probability of $E$ in every subpopulation of $p$.\(^7\)

Sober and Wilson illustrate Simpson’s Paradox with an example of a discrimination inquiry at the University of California Berkeley.\(^8\) Based on the smaller overall percentage of women who were admitted than that of men, it was suggested

---


\(^8\) Nancy Cartwright presented this example to illustrate Simpson’s Paradox in her “Causal Laws and Effective Strategies,” Special Issue on Counterfactuals and Laws, *Nous* 13, no. 4 (November 1979): 419-437. In that article she presents another illustrative example involving smoking and heart disease. There is a correlation between these: smoking causes heart disease. Given this, we would expect there to be a higher probability of heart disease in smokers than in non-smokers. Cartwright says that this doesn’t necessarily follow. She suggests that a statistical anomaly where the probability of heart disease declines in smokers could obtain if smoking was also correlated with something like exercise, which reduces the probability of heart disease (so long as that reduction occurs at a higher rate than smoking increases one’s chance of heart disease).
that the University’s admissions policies were discriminatory. Upon further investigation, however, it was discovered that every department admitted an equal proportion of women to men. And yet fewer women, overall, were admitted than men.

This seemingly paradoxical result can be explained as follows. It was discovered that a greater number of women tended to apply to departments that had lower acceptance rates than those to which men tended to apply.\(^9\) Let us say that department A accepts only 25% of its applicants, while department B accepts 75%. Suppose that department A receives the following distribution of applicants: 80 women and 20 men. Department A accepts 25% of the women applicants and 25% of the men applicants: 20 women and 5 men. Suppose now that 20 women and 80 men apply to department B. Since department B accepts 75% of their applicants, the result is that 15 women and 60 men are accepted.

A combined total of 100 men and 100 women applied to the two departments. Only 35 women in total were admitted, while 65 men were admitted. And yet, no department was discriminatory\(^{10}\) in their acceptance rates, since each department accepted an equal proportion of men to women. Thus, one way that Simpson’s Paradox is thought to be resolved is by appreciating the underlying causes that the

\(^9\) According to Nancy Cartwright (1979) and Judea Pearl (2000), it is this correlated cause that is responsible for the statistical anomaly. Pearl says: “In the case of Simpson’s paradox, we have a clash (i) between the assumption that causal relationships are governed by laws of probability calculus and (ii) the set of implicit assumptions that drive our causal intuitions (7).”

\(^{10}\) What Sober and Wilson mean here is that no department was violating proportionality. But some might say that proportionally admitting men and women is discriminatory. One might, for example, argue that, one must admit an equal number of men and women in order to be non-discriminatory.
reversal reveals. That is, in the Berkeley case, the underlying explanation for the phenomenon was that there was a correlation between being female and applying to programs whose acceptances rates were lower.

According to Sober and Wilson, the existence of altruism is an example of Simpson’s Paradox:

A bias exists in the two departments combined, despite the fact that it does not exist in any single department, because the departments contribute unequally to the total number of applicants who are accepted. In just the same way, altruists can increase in frequency in the two groups combined, despite the fact that they decrease in frequency within each group, because the groups contribute unequally to the total number of offspring.\(^{11}\)

As I understand them, Sober and Wilson use Simpson’s Paradox to reveal that it doesn’t necessarily follow that, based on the fact that altruists will be less fit in a subgroup than selfish individuals, they will necessarily be less fit overall.\(^{12}\) While


\(^{12}\) There is some controversy over how to interpret Simpson’s Paradox and also whether Sober and Wilson employ it appropriately for their purposes. A fuller discussion of these issues lies beyond the scope of my aim here. According to Pearl, what Simpson’s reversals reveal is that there is some kind of correlated 3rd cause present that yields the observed result. It’s not clear that Sober and Wilson recognize this. Peter Gildenhuys says:

Sober and Wilson have misunderstood how Simpson’s paradox operates….Simpson’s paradox is not really a paradox since there is a clear explanation for what is going on in such scenarios. A purported cause, being a man, does not lead to its purported effect, getting accepted at Berkeley, because of the operation of another cause that works to skew the numbers. In the Berkeley case, the alternate cause in operation is applying to easy departments. It turns out that being a man is correlated with applying to easy departments and being a woman is correlated with applying to tough departments. That is what explains the overall bias in the acceptance of candidates to the graduate school. Sober and Wilson
selfish individuals may do better within groups than altruists, if two groups are combined, the reverse effect can be obtained. Thus, “what is true, by definition, is that altruists are less fit than selfish individuals in the same group…however, nothing follows from this as to whether altruists have lower fitness when one averages across all groups…. [G]roup and individual selection are opposing forces; which force is stronger determines whether altruism increases or declines in frequency in the ensemble of groups. Just as Darwin conjectured, it takes group selection for altruism to evolve.” And further, “if the process is repeated over many generations, altruists will gradually replace the selfish types, just as the selfish types replaced the altruists in the one-group example.”

We thus get a sketch of the kind of contribution group selection can make to explaining the emergence of traits that appear to perform to the detriment of their possessors and to the advantage of the group in which they occur. Individual explicitly assume genetic determinism in their model, so the only traits that make a difference to the evolution of altruism in their model are the genes for altruism and selfishness along with the two traits, being among mostly nonaltruists, and being grouped with the majority of altruists, that are acquired when the population divides into subgroups. The distribution of these last two traits is what is responsible for the success of the altruists over the nonaltruists. Being surrounded by altruists increases fitness, and altruism is correlated with being surrounded by altruists while selfishness is correlated with being surrounded by nonaltruists. “The Evolution of Altruism: The Sober/Wilson Model,” Philosophy of Science 70 (January 2003): 36.


14 Sober and Wilson, Unto Others, 26.

15 Sober and Wilson point out that:

[G]roup selection favors any behavior that increases the relative fitness of groups. Altruism fits this definition, but it also has the additional feature of decreasing the relative fitness of individuals within groups. In other words, evolutionary altruism is a two-dimensional concept that includes both benefits to others and costs to self. A trait that increases group fitness
selection may favor the evolution of selfishness within groups (and thus altruists will be less fit relative to non-altruists within a single group), but members of altruistic groups will do better than members of selfish groups. As Sober and Wilson say, “individual selection favors traits that maximize relative fitness within single groups. Group selection favors traits that maximize the relative fitness of groups. Altruism is maladaptive with respect to individual selection but adaptive with respect to group selection. Altruism can evolve if the process of group selection is sufficiently strong.”

The idea that natural selection can operate at the level of groups as well as that of individuals is not new, and was invoked by Darwin himself to explain human moral behaviour:

It must not be forgotten that although a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe, yet that an increase in the number of well-endowed men and advancement in the standard of morality will certainly give an immense advantage to one tribe over another.\(^{17}\)

\(^{16}\) Sober and Wilson, *Unto Others*, 27.

But while the concept of selection acting at the level of groups and the promise to explain the emergence of genuinely altruistic behaviours is attractive, the plausibility of the theory is highly contentious. There are two lines of criticism in particular that I will take up. The first raises objections specific to Sober and Wilson’s articulation of altruism, and to their endorsement of multi-level selection and conception of a group, and calls into question the success of their view in meeting their aims. The second (and I think most pressing) objection is that the requirements for biological group selection in the human case are unlikely and, consequently, it is unlikely that group selection played a significant role in the evolution of altruism in humans. I will take these in turn.

**Internal Difficulties**

*Genuine versus Apparent Altruism*

We have seen that one motivating factor behind Sober and Wilson’s endorsement of group selection is to provide an account of *genuine* (as opposed to merely *apparent*) altruism. They are careful to draw the distinction between what they take to be cases of genuine altruism, which involve genuine fitness costs to the individual performing them, and cases of apparent altruism, where the apparent

---

18 George C. Williams’s *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought* (Princeton: Princeton University Press, 1966) called into question the theoretical plausibility of group selection by claiming that natural selection occurs at the level of genes, and that this rules out the possibility of selection at the level of groups. Sober and Wilson think otherwise, and attempt to breathe new life into the theory.
fitness cost is recouped at the individual level. Sober and Wilson claim to show that genuine altruism exists, but seem to be able to do so only by adopting the special “token” form of altruism, according to which the altruistic nature of an act is to be determined by that particular act. One might suggest, however, that the same “recouping” occurs in the cases Sober and Wilson cite as genuine cases of altruism, but instead at the level of the group. In other words, if their group selection really works, it also makes altruism only an appearance. The cost of altruism is explained away in terms of the benefit conferred onto the group instead of the individual. In order for altruism to evolve, it must be selectively advantageous at whatever level, group or individual. If so, then Sober and Wilson have succumbed to their own criticism. And if we resist this conclusion, and contend that recouping the fitness loss at the group does not eliminate altruism in the same way as it does at the individual level, then we need to ask whether Sober and Wilson are pulling a semantic sleight of hand to obtain their desired result. It is not clear why recouping costs at an individual level should disqualify an act from being genuinely altruistic, but recouping them at a group level should not.

There is an interesting parallel here between Sober and Wilson’s analysis of altruism and Gauthier’s analysis of the rationality of actions, which we will examine in chapter five. On Gauthier’s view (which I will defend), acts that are disadvantageous when taken as discrete acts can be rational if they are part of a strategy that is, on the whole, beneficial. If we take this parallel seriously, then, just as

---

19 Sober and Wilson, *Unto Others*, 31-35.
it is not irrational to perform disadvantageous acts so long as they are part of a
disposition that it is rational to have, so too, we might conclude, it is not genuinely
altruistic to perform actions that are at a selective disadvantage within groups, but
which are advantageous at the between-group level. What appears to be irrational in
the first case turns out not to be in the context in which we find it. Likewise, what
appears to be altruistic in the second case turns out not to be in a parallel context. The
implications of this parallel in the biological context do not make much of a material
difference, insofar as the central aim there is to uncover the underlying processes
responsible for and subsequent explanation of the observed behaviour in a way that is
compatible with evolution. However, the implications are more significant if one’s
aim is, like Sober and Wilson’s, to show that some seemingly altruistic behaviour is
genuine, for it suggests that one of their central motivations for adopting a group
selection model cannot be met.

*Multi-Level Selection and the Concept of a Group*

The second problem for Sober and Wilson comes from combining their view
of multi-level selection with their view of what counts as a group. I will first lay out
their views on these matters and then argue that they do not combine well.

Sober and Wilson adopt a multi-level selection view, which maintains that
evolutionary processes operate at different biological levels. This lies in contrast to
the individual-level approach, which explains the emergence of traits solely in terms
of individual selection pressures. On this view, if altruism is to evolve, it must do
better, on average, than other strategies within a population. The multi-level perspective, by contrast, will maintain that altruism can be at a selective disadvantage within groups, but increase in numbers in the global population.

Sober and Wilson think that, in adopting this multi-level selection framework, we get a richer picture of the evolutionary mechanisms in operation. They say:

[A] gene can evolve by increasing its fitness relative to other genes within the same individual, by increasing the fitness of the individual relative to other individuals within a group, or by increasing the fitness of the group, relative to other groups in the total population. The net effect of these nested levels of selection determines the relative fitness of alternative alleles in the total population. [Multi-level selection] correctly predicts what evolves (a fact that is not in dispute) and has the virtue of treating every level of the biological hierarchy in the same way.

There are two primary ways in which we can characterize the processes involved in a multi-level selection framework. We can hold that there are a number of different perspectives from which to view the evolution of particular traits, but that no single one perspective captures any essential characteristic of the process better than any other. On this view, sharp teeth in tigers can be explained as a result of either (i) competition of sharp-toothed tigers against dull-toothed tigers or (ii) genes coding for

---

20 Sober and Wilson, Unto Others, 32.

sharp teeth competing against genes coding for dull teeth, but neither (i) nor (ii) gives a fuller explanation of the emergence of sharp teeth among tigers.

Alternatively, we might maintain that a proper explanation of particular traits requires an appeal to multiple processes operating at different levels. On this view, for example, altruism is to be regarded as a result of the processes of individual selection operating at the within-group level and group selection operating at the between-group level. While an individual-level perspective (according to which altruism is a result of altruists having a higher average fitness in the global population) might yield an accurate prediction of the concentration of altruists that will evolve, it will not provide an accurate account of the causes underlying the evolution of altruism.22

Sober and Wilson endorse this second perspective, referred to by Kim Sterelny and Philip Kitcher as “hierarchical monism.”23 Sober and Wilson say:

A predictive theory needs to focus on the variables, which are the fitness differences that can occur anywhere in the biological hierarchy. Multilevel selection theory offers a precise framework for identifying these differences (among genes/within individuals, among individuals/within groups, among groups/within global populations, etc.) and for measuring their relative strengths. Selfish gene theory requires these same distinctions, but its central concept of genes as replicators offers no help. All the hard work is left for the

---


tag-along concept of vehicles, which is not nearly as well developed as multilevel selection theory.\textsuperscript{24}

According to this view, while an individual-level characterization of the evolution of traits such as altruism will yield correct predictions, a multi-level selection perspective that includes group selection will provide us with greater insight into the nature of the evolution of altruistic behaviour.

When altruism evolves, there typically are two processes at work. Between-group selection favors the evolution of altruism; within-group selection favors the evolution of selfishness. These two processes oppose each other. If altruism manages to evolve, this indicates that the group-selection process has been strong enough to overwhelm the force pushing in the opposite direction. When this two-level process occurs in a population, an appropriate causal analysis should describe what is going on. The summary statement that the trait that evolved had the higher average fitness does not include any of these details. It is neutral on what process or processes are responsible for that result. When one trait is more fit than another, this may be due to pure individual selection, to pure group selection, or to a mixture of the two. The description of the effect fails to specify what the causes were.\textsuperscript{25}

I will now argue that Sober and Wilson’s adherence to this multi-level perspective raises internal difficulties for their view when combined with their conception of what makes a group.

\textsuperscript{24} Sober and Wilson, \textit{Unto Others}, 91-92.

\textsuperscript{25} Sober and Wilson, \textit{Unto Others}, 33.
A group, according to Sober and Wilson, “is defined as a set of individuals that influence each other’s fitness with respect to a certain trait but not the fitness of those outside the group. Mathematically, the groups are represented by a frequency of a certain trait, and fitnesses are a function of this frequency. Any group that satisfies this criterion qualifies as a group in multilevel selection theory, regardless of how long it lasts or the specific manner in which groups compete with other groups.”

Thus, on their view, “groups are defined exclusively in terms of fitness effects and everything else about groups, such as their duration and the manner in which they compete with other groups, follows from the nature of the interaction.”

On this view, kin selection and reciprocal altruism (which are more commonly classified as instances of individual selection) both emerge as instances of group selection. In the case of kin selection, the group happens to be related to one another; in cases of reciprocal altruism, group size may be as small as two. What qualifies these types of interactions as group interactions are the fitness consequences of these interactions. Sober and Wilson say that “when stated as a group-selection process…several interesting features emerge….Hamilton’s inclusive fitness theory is basically a different model of analyzing the same process. It correctly predicts the final outcome but does not distinguish clearly between the opposing forces of group and individual selection, as they are defined within the group selection tradition.”

---

26 Sober and Wilson, Unto Others, 93.
27 Sober and Wilson, Unto Others, 93.
28 Sober and Wilson, Unto Others, 92-98.
Sober and Wilson contend that this broader conception of a group allows them to point to instances of apparent altruism, such as what is displayed among relatives, and claim that these are instances of genuine altruism that arise through group selection. This re-orientation permits Sober and Wilson to claim that group selection is a prevalent force in nature. It is, however, not clear what utility this serves. On this, Robert Trivers says:

Sober and Wilson's real examples of group selection are examples of group selection under highly particular...conditions. The rest of their effort is an attempt to convince us that group selection is a globally important process and, indeed, the major selective factor molding all of human behavior. They can make this claim only by reinterpreting kin-selection as a form of group selection, treating reciprocal altruism as a form of group selection, calling all instances of social interaction between individuals, however fleeting or few their number, examples of groups and hence acted upon by group selection, and then by failing to explore in a rational way, or indeed at all, the power of group selection—given even low levels of between-group migration—compared to the power of individual or genetic selection.  

Nor is it clear that this conception yields desirable implications. Samir Okasha says that “to argue that game-theoretic and ESS [evolutionary stable strategy] models are actually versions of group selection is more than a bit odd. The logic behind Sober

29 Sober and Wilson, *Unto Others*, 177.

and Wilson’s argument is clear, but it leads to strange places.”31 One such strange place, identified by Matthew Barrett and Peter Godfrey-Smith, is that this looser conception of a group according to which groups are everywhere presents a tension with Sober and Wilson’s endorsement of hierarchical monism. Recall that, on this view, group-level explanations must be included wherever they are present. And since a broad conception of a group permits a lot to qualify as such, group selection crops up everywhere.

On its own, this implication is not problematic. But the fact that examples of group selection are consequently abundant seems to diminish the needed contrast between causal forces in order to support Sober and Wilson’s contention that group selection is needed to provide a causally accurate representation of the forces of natural selection at work. As Barrett and Godfrey-Smith put it:

Sober and Wilson’s hierarchical monism, and their ultra-liberal conception of groups, are a bad match. Ultra-liberalism means that group selectionist explanations are almost always available—wherever inter-organismal interactions affect fitness. Hierarchical monism means that where they are available, they are mandatory in causal explanation. As a consequence, too much counts as group selection.32

Barrett and Godfrey-Smith point out that “the problem here arises from the combination of ultra-liberalism about groups and hierarchical monism. The problem


32 Barrett and Peter Godfrey-Smith, 688.
does not arise—or arises far more benignly—for those who combine a liberal conception of groups with a stronger form of pluralism, as they do not preclude an alternative description in individualist terms.”

If so, the above difficulty may be able to be resolved, but that would mean for Sober and Wilson to give up at least one component of their view. It might, however, be the case that Sober and Wilson will not see this as a difficulty that needs resolving and be happy to conclude that the causal processes underlying group selection are nearly omni-present. But even so, as we will see in the next section, there remains a difficulty with respect to the possibility of biological group selection playing a significant role in the evolution of altruism in the case of humans.

**Empirical Plausibility**

We have seen that group selection is a general framework that promises to explain the evolution of traits that are at a selective disadvantage at the individual level but that permit members of the group where those traits are found to outcompete members of groups where those traits are not found. We have also seen that Sober and Wilson count a lot more as a group than is normally done. On their view, all of the examples of the emergence of altruism through individual selection models like kin selection and reciprocity are actually models of group selection. If we grant to Sober and Wilson their looser conception of a group, then, insofar as the conditions required for altruism to emerge in cases involving interactions between kin and those

---

33 Barrett and Godfrey-Smith, 688.
governed by reciprocity can obtain, there will be examples of the emergence of altruism by group selection. But even if we say that those are clear examples of group selection, it does not follow that group selection gives a plausible explanation of widespread cooperation in human societies. Kin selection and reciprocal altruism, we have seen, are commonly employed and widely accepted as mechanisms underlying certain examples of altruistic behaviour. But whether we refer to the evolution of these instances of altruism as a result of group selection does not fill the explanatory gap that the distinctively high scope and scale of human cooperation produces. I do not deny that kin selection and reciprocity are important explanatory mechanisms, but they are not sufficient to explain cooperation in the human case.

Sober and Wilson maintain that the evolution of altruism is an outcome of a conflict of two competing processes: individual selection within groups, and group selection between groups. For altruism to evolve by group selection, altruists must be distributed in populations in such a way that groups containing altruists will increase at a more rapid rate than groups that contain selfish individuals and, thus, that the global population of altruists will grow in spite of their being disadvantaged at the individual level. Maynard Smith produced the Haystack model,\textsuperscript{34} which modeled the specific conditions required for altruism to evolve. They are as follows. (1) Groups must be isolated and sufficiently varied; otherwise the effects of group selection will be negated. (2) Groups must divide and intermix at just the right time to permit

\textsuperscript{34} Maynard Smith, “Group Selection and Kin Selection,” \textit{Nature} 201 (1964): 1145-1146. The Haystack model supposes that each haystack contains two mice. In the summer, these mice reproduce asexually. In the fall, the haystacks are removed by farmers. At this point, the mice retreat to the meadow. They settle randomly, two mice in each, once new haystacks are built.
differential reproduction of altruism and selfishness and also to prevent individual selection from driving altruism to extinction. (3) Groups must then re-isolate themselves, and the process must then be repeated. Maynard Smith was skeptical that these conditions would actually obtain in nature, and thus that group selection played much of a role in the evolution of altruistic tendencies.

But while these conditions may plausibly be met when “groups” are small or composed of kin, larger human societies do not naturally compose themselves in this way. As Richerson and Boyd say, “the trouble with a straightforward group selection hypothesis is our mating system. We do not build up the concentrations of intrademic relatedness like social insects, and few demic boundaries are without considerable intermarriage.” And further, “even very small amounts of migration are sufficient to reduce the genetic variation between groups to such a low level that

---

35 Sober and Wilson outline what is required for group selection to occur as follows:

First, there must be more than one group; there must be a population of groups. Second, the groups must vary in their proportion of altruistic types. Third, there must be a direct relationship between the proportion of altruists in the group and the group’s output; groups with altruists must be more fit (produce more offspring) than groups without altruists. Fourth, although the groups are isolated from each other by definition (the S types in group 1 do not benefit from the A types in group 2), there must also be a sense in which they are not isolated (the progeny of both groups must mix or otherwise compete in the formation of new groups). These are the necessary conditions for altruism to evolve in the multigroup model. To be sufficient, the differential fitness of groups (the force favoring the altruists) must be strong enough to counter the differential fitness of individuals within groups (the force favoring the selfish types). Unto Others, 26.

36 In such cases, there are other factors that help to prevent “subversion from within”—that is, from selfish individuals driving altruists to extinction within the group before the benefit to the group can be obtained. In the case of 2-person interactions, variation between groups and similarity within is easily met. And in the case of kin groups, these conditions can also easily be met, and are supplemented by the influence of genetic relatedness.

group selection is not important.”38 If so, to the extent to which similarity within and
variation between biological groups is needed for altruism to evolve, biological group
selection is unlikely to have played a significant role in the evolution of altruism in
humans.

Thus, even if we grant to Sober and Wilson that kin selection and reciprocity
are examples of group selection, the interesting processes for our purposes—viz.,
explaining widespread cooperation among non-related individuals—still haven’t been
addressed. To explain the origins of human cooperation we need more than just a
rewriting of terminology. We need to be able to show that something like what
happens in Maynard Smith’s haystack example could actually have occurred in
human populations. And this is unlikely.

**Cultural Group Selection**

But while group selection at the biological level is unlikely to have had much
effect in producing altruistic individuals due to the implausibility that genetic
differences will be distributed in such a way that there is sufficient variation between
groups and sufficient similarity within groups, this does not hold in the case of
cultural variants. The most notable account of the role of cultural group selection in
the evolution of altruism is found in the works of Peter Richerson and Robert Boyd,
to which we will turn in the next chapter. They contend that cultural variants in the
form of social norms will help to generate suitable variation between groups and

---

homogeneity within groups to permit group selection at the cultural level. Some groups will operate with more successful norms than others and, thus, will outcompete groups operating with less successful norms, and as those groups flourish so will their norms.

Sober and Wilson recognize the role of culture—especially cultural norms—in potentiating group selection in the human case, but they maintain that “cultural evolution per se is not more favorable than genetic evolution for group selection.” According to them, not all group selection is culturally driven, but natural group variations can be changed or amplified through cultural norms in the case of humans. Thus, “social norms and cultural evolution are not required for group selection among alternative stable equilibria, but they vastly increase its potential.”

But if what I have argued is correct, Sober and Wilson overstate the significance of biological group selection in the evolution of altruism among human societies. It is unlikely that the genetic variability exists naturally in larger human groups to the extent that is needed to get biological group selection of altruism to occur in those cases. I will, in the next chapter, turn to Richerson and Boyd’s account

---

39 This is dealt with in Unto Others, 149-158. There Sober and Wilson say, for example:

Human social groups are never genetically uniform, but they are often quite uniform behaviorally, especially when the behaviors are reinforced by social norms. A new behavior does not necessarily spend a long time at low frequency—it can quickly become the only behavior that is practiced by a social group. Heritability can be even greater for groups than for individuals when cultural processes are involved. A social group can maintain its distinctive behavioral characteristics even though the members of the group are constantly changing (150).

40 Sober and Wilson, Unto Others, 150.

41 Sober and Wilson, Unto Others, 152.
of cultural group selection and argue that this account promises to provide a comprehensive account of the evolution of human cooperation. We will turn to a discussion of the relevance of this to normative ethics in the final chapter.
CHAPTER FOUR
The Significance of Culture

In the last chapter we looked at Sober and Wilson’s account of the evolution of altruism in terms of biological group selection. I argued that the likelihood that biological group selection played a significant role in the evolution of human cooperation is low, and that this is largely due to the requirement of variation between groups and homogeneity within groups, and the implausibility that these requirements can be met at the biological level among larger human groups. I suggested that cultural group selection, as acknowledged to some extent by Sober and Wilson, and developed by Richerson and Boyd, promises to resolve deficiencies of earlier non-culturally-centered accounts in explaining the origins of human cooperation. In what follows I will outline the central elements of Richerson and Boyd’s view, and will argue that it does indeed help to explain the origins of human cooperation in a way that past accounts could not.

Setting the Stage for Group Selection

Recall that what is required for group selection to occur is (1) there must be sufficient variation between groups, (2) there must be sufficient similarity within groups, and (3) there must be an adequate balance between selection at the individual
level and group level so that the outcome of selection is not the same as whatever is arrived at via individual selection. In other words, individual selection cannot be so strong as to render negligible any selection at the group level. We saw in the last chapter that within-group selection must be weak enough in order to ensure that selfish individuals do not take over populations before they intermix with other populations to permit the effects of group selection. I argued that what makes the evolution of altruism by biological group selection implausible is that it is unlikely that the arrangement of groups will be such that individual-level selection won’t outpace group selection. More specifically, what is missing from the biological account is the plausibility that genetic differences will be distributed in such a way that there is sufficient variation between groups and sufficient similarity within groups. According to Richerson and Boyd, what is missing is culture.

Richerson and Boyd define “culture” as “information capable of affecting individuals’ behavior that they acquire from other members of their species through teaching, imitation, and other forms of social transmission.”¹ They go on to explain that:

By information we mean any kind of mental state, conscious or not, that is acquired or modified by social learning and affects behavior. We will use everyday words like idea, knowledge, belief, value, skill, and attitude to describe this information, but we do not mean that such socially acquired

information is always consciously available, or that it necessarily corresponds to folk-psychological categories. Our definition is rooted in the conviction that most cultural variation is caused by information stored in human brains—information that got into those brains by learning from others.²

Richerson and Boyd think that culture increases human adaptability. They say:

Culture has made the human species a spectacular ecological success. Since the first appearance of tools and other evidences of culture in the archaeological record, the human species has expanded its range from part of Africa to the entire world, increased in numbers by many orders of magnitude, exterminated competitors and prey species, and radically altered the earth’s biota.³

Rapid changes in the environment of our ancestors likely favored a way of economically adjusting to changing conditions. This is the role that culture played. Cultural evolution refers to the process by which cultural information is transmitted. Just as genes are transmitted through genetic evolution, so is culture transmitted through cultural evolution. Cultural evolution works much more quickly than biological evolution and allows humans to adapt quickly to environmental or social changes. The chief mechanism for cultural evolution is social learning, mechanisms for which include teaching, imitation, and cognitive biases.

² Richerson and Boyd, Not by Genes Alone, 5.

Social learning involves learning from others, and is to be distinguished from individual learning. While individual learning permits individuals, primarily through trial-and-error, to adapt to varying environments, such learning can often be time consuming and error prone. Social learning permits one to avoid these costs. Consequently, humans have evolved capacities for social learning, such as the ability to learn from others through teaching and imitation. Social learning also makes possible cumulative cultural evolution, which amounts to “behaviors or artifacts that are transmitted and modified over many generations, leading to complex artifacts and behaviors.”

However, social learning can sometimes be inaccurate. Imitation and learning must be selective in order to ensure that one is learning from individuals whose behaviour is best suited to the environment. A rapidly changing environment might, for example, result in some behaviours that are best responses to a particular environment being adopted via imitation in an environment to which they are no longer best suited. Accordingly, human psychology has evolved particular biases that guide the kind of information that is adopted.

[C]ultural transmission is biased by decision rules that individuals apply to the variants they observe or try out. The rules behind such selective imitation may be innate or the result of earlier imitation or a mixture of both. Many types of rules might be used to bias imitation. Individuals may try out a behavior and let reinforcement guide acceptance or rejection, or they may use various rules of thumb to reduce the need for costly trials and punishing errors. Rules like

---

“copy the successful,” “copy the prestigious” or “copy the majority” allow individuals to acquire rapidly and efficiently adaptive behavior across a wide range of circumstances, and play an important role in our hypothesis about the origins of cooperative tendencies in human behavior.⁵

Richerson and Boyd suggest that symbolic marking may also have evolved to help individuals identify in-group members.⁶ In-group members are more likely to be well-adapted to their environment than foreigners and, thus, individuals are better off learning from fellow group members. Identification of in-group members increases the chances of learning how to live successfully in the environment, and of successful social interaction. Symbolic markers delineate groups of individuals into distinct ethnic groups, and include such things as dress, speech, rituals, and so on. The presence of symbolic markers allows individuals to identify group members, and thus facilitates selective imitation and social interactions. As Richerson and Boyd say: “Rapid cultural adaptation makes the local population a valuable source of information about what is adaptive in the local environment. Individuals are well advised to imitate locals and avoid learning from immigrants who bring ideas adapted to other environments.”⁷ And Richerson and Boyd claim that, “once there are reliable

---


⁶ Richerson and Boyd, Not by Genes Alone, 211-213.

⁷ Boyd and Richerson, The Origin and Evolution of Cultures, 100.
symbolic markers, selection will favor the psychological propensity to imitate and interact selectively with individuals who share those markers.”

It is important to point out that there is evidence that non-human animals can learn socially and, thus, possess culture. But what distinguishes human cultures from those of non-humans is the way in which culture is transmitted. This covers not only the heuristics and biases by which we gather information, but also the ability to build on the discoveries of others such that cumulative cultural evolution is possible. Richerson and Boyd report that selective imitation and cumulative cultural evolution are unique to human beings, and they locate this divergence in our unique capacity for imitation. Richerson and Boyd distinguish between social enhancement and imitation. Social enhancement occurs when individuals adopt particular behaviours or skills as a result of their particular surroundings. Some monkeys, for example, will learn to use rocks to smash nuts, not because they observe others doing so, but because they learn to do so individually. Imitation, by contrast, permits the acquisition of skills without having to learn them on one’s own. According to Richerson and Boyd, “all known cases of animal social traditions can be explained as the result of social enhancement.” It follows that we need not invoke learning by imitation in order to explain these animal behaviours. For Richerson and Boyd, imitation is required for both selective imitation and cumulative cultural evolution.

---

8 Boyd and Richerson, *The Origin and Evolution of Cultures*, 100.

9 Boyd and Richerson, *The Origin and Evolution of Cultures*, 44.

Thus, “understanding the evolution of the psychological mechanisms that allow imitation is key to understanding human evolution.”

**Group Stability and Between-Group Differences**

We are now in a position to see how cultural evolution makes the emergence of cooperation possible. Cultural adaptation has resulted in significant behavioural differences between groups. Culture permits individuals to adapt quickly to environmental changes. Consequently, individuals are very locally adapted to a wide range of environments, which has resulted in significant variations in behaviours between human groups. But the existence of this variation is not sufficient to permit group selection. Culture can be readily transmitted between groups, primarily through migration, which, on its own—viz., in the absence of any countervailing mechanisms—is liable to reduce the variation between groups. And if the variation between groups is minimal, group selection becomes unlikely. Thus, if group selection on cultural variants is to be a significant force, there must be a mechanism in place to ensure that this variation is maintained, i.e., that the groups are stable.

This stability is provided by a conformist bias aided by in-group marking. Richerson and Boyd say:

The conformist effect overcomes the critical problem with group selection. In the case of a genetic system of inheritance, variation between groups tends to evaporate quickly in the face of modest amounts of migration. In the case of

---

altruistic traits, selection within groups against altruists also reduces between-group variation for altruism. The existence of large-scale cooperation in human societies invites a group-functional interpretation, and perhaps the peculiarities of the cultural system of inheritance are responsible.¹²

Conformity—and thus group stability—is maintained by the fact that (as we have seen) human beings are imitative, and copy those around them such as parents, peers, successful people, and particularly what the majority of people do. If individuals are copying whatever the majority is doing, then the more popular norms will begin to be more common. This in turn will cause the less popular norms to become even rarer, presumably to the point where the more popular norms take over the population entirely. And, “if the effect is strong compared with migration, then variation among groups can be maintained.”¹³ As Adrian Bell says:

An outcome of some social learning biases is the maintenance of between-group differences in the face of heavy migration. While migration between groups can eliminate between-group genetic variation rather quickly, some social learning biases may maintain high levels of between-group cultural variation despite it. Prestige, conformity and guided variation, are important candidate learning preferences for this reason.¹⁴

¹² Richerson and Boyd, Not by Genes Alone, 163.
¹³ Richerson and Boyd, Not by Genes Alone, 205.
Variation between groups can also be maintained by moralistic punishment. Moralistic punishment consists in visiting social sanctions on those who fail to conform to the group’s norms. This may take the form of attacking, shunning, or gossiping about them, denying them access to territories or mates, or kicking them out of the group entirely, and works to generate a homogeneity within groups. In a population that adheres to norm X, for example, members might punish any member who maintains any norm other than X.

Richerson and Boyd contend that moralistic punishment of defectors provides a greater stability to cooperation than does reciprocal altruism.

Moralistic punishment is more effective in supporting large-scale cooperation than reciprocity for two reasons. First, punishment can be targeted, meaning that defectors can be penalized without generating the cascade of defection that follows when reciprocators refuse to cooperate with defectors. Second, with reciprocity, the severity of the sanction is limited by the effect of a single individual’s continued cooperation on each other group member, an effect that decreases as group size increases. Moralistic sanctions can be much more costly to defectors, so that cooperators can induce others to cooperate in large groups even when they are rare. Cowards, deserters, and cheaters may be attacked by their erstwhile compatriots, shunned by their society, made the targets of gossip, or denied access to territories or mates. Thus, moralistic punishment provides a much more plausible mechanism for the maintenance of large-scale cooperation than reciprocity.¹⁶

¹⁵ Richerson and Boyd, Not by Genes Alone, 200. Examples of this kind of punishment are the British placement of criminals in Australia and the United States’s placement of them in Alcatraz.

¹⁶ Richerson and Boyd, Not by Genes Alone, 200.
They say further:

If moralistic punishment is common, and punishments sufficiently severe, then cooperation will pay. Most people may go through life without having to punish very much, which in turn means that a predisposition to punish may be cheap compared with a disposition to cooperate (in the absence of cooperation). Thus, relatively weak evolutionary forces can maintain a moralistic predisposition, and then punishment can maintain group-beneficial behavior.17

Thus, human beings have a strong conformative bias, and this, together with a tendency to punish those who do not think like them on basic social norms, provides a homogeneity to culture and makes variations in culture stable. Those who join groups tend to accept the norms, and those who reject the norms tend to migrate to other groups. This transmission rule reduces variation within populations and maintains between-population differences. This stability of variation makes cultural group selection possible in a way that genetic group selection is not.

**Group Selection and the Emergence of Cooperation**

Societies can be operating with a number of different cultural variants (or norms). Some groups might boil drinking water; others might chlorinate it; still others might drink it straight. Some of these norms might be better than others; sometimes

---

17 Richerson and Boyd, *Not by Genes Alone*, 201.
no one norm is better than another (as in driving on the left or right side of the road—so long as everyone else is doing the same thing).

Group selection results in the selection of group-beneficial traits. Groups that operate with more advantageous norms will outcompete those groups that operate with less advantageous norms. Consider, for example, that Group A boils drinking water; B does not. B dies of cholera. Boiling water is thus the cultural variant that spreads. Richerson and Boyd point out that:

Winning groups must replace losing groups, but losers need not be killed. The members of losing groups just have to disperse or be assimilated into the victorious group. If losers are resocialized by conformity or punishment, even very high rates of physical migration need not result in the erosion of cultural differences.\(^\text{18}\)

Thus when groups with different cultures with differential fitnesses come into conflict, and one wins out over the other, the culture of the winning one grows, as the losing group either is extinguished or absorbed. On their own, the norms of the societies that lose out might permit societies operating with them to function satisfactorily. But those norms may not serve societies well when the societies are in competition with other societies in the world.

Richerson and Boyd, following Darwin, think that cooperative groups will tend to outcompete non-cooperative groups and, thus, a cooperative culture will take root and grow. Prisoner’s Dilemmas show us that each individual’s best course of

\(^{18}\text{Richerson and Boyd, Not by Genes Alone, 207.}\)
action is to defect. But we see that if each individual does that (viz., performs the “non-cooperative” action), then each individual will end up worse off than they would have had they both cooperated. Let us now extend this. Suppose that we have two groups, each composed of two players engaged in a Prisoner’s Dilemma situation. Utility payoffs are as follows: mutual cooperation will yield a payoff of 3 for each player, mutual defection a payoff of 2, and unilateral defection will yield a payoff of 4 for the defector and 0 for the cooperator. Suppose that the individuals in Group A are rational utility-maximizers (who lack cooperative dispositions). Each individual, being rational, will defect. Each will then get a payoff of 2. Suppose now that the members of Group B are not rational utility-maximizers. That is, they are predisposed to cooperate in such situations. Each individual will then cooperate, resulting in a payoff of 3 each. If we translate these payoffs to numbers of offspring, then we will see that rational utility-maximizers will generate 2 offspring, while cooperative individuals will produce 3. Thus, “Since larger, more-cooperative, and more-coherent groups should outcompete smaller, less cooperative groups,” Richerson and Boyd conclude that “group selection could give rise to culturally transmitted cooperative, group-oriented norms, and systems of rewards and punishments to ensure that such norms are obeyed.”

Richerson and Boyd say that “intergroup competition is not the only mechanism that can lead to the spread of group-beneficial cultural variants.”

---


Group-beneficial variants can also spread through imitation of one’s neighbours. If neighbouring groups are observed employing successful cultural variants, it is possible, they claim, that other groups might adopt these through imitation. The spread of Christianity in the Roman Empire, they suggest, can be explained by this process.\(^21\)

**Coevolution of Genes and Culture**

Thus, cultural evolution generates differences between groups. These differences make for an environment conducive to group selection. This is the principal contribution of cultural evolution to the evolution of sociality. Richerson and Boyd point out:

> In other primate species there is little heritable variation among groups within a species. The behavior of groups depends on the habitat and ecology, the demographic structure, and the personalities of particular individuals. But these differences are small and ephemeral, and, as a consequence, group selection at the level of whole primate groups is not an important evolutionary force. In contrast, it is an empirical fact that there is much heritable cultural variation among human groups….This suggested to us that group selection might be a more important process shaping human behavior than the behavior of other animals.\(^22\)

---

\(^21\) Richerson and Boyd, *Not by Genes Alone*, 209.

\(^22\) Boyd and Richerson, *The Origin and Evolution of Cultures*, 134.
More cooperative groups do better than less cooperative groups. The resulting environment shaped by group selection in turn favors prosocial psychological mechanisms. As Richerson and Boyd put it, “if generally cooperative behavior is favored in most social environments, selection may favor genetically transmitted social instincts that predispose people to cooperate and identify within larger social groupings.”

For example, “selection might favor feelings such as guilt that make defection intrinsically costly, because this would bring the costs of defection into the present, where they would be properly compared with the cost of cooperation.”

In order to draw this conclusion regarding the emergence of prosocial dispositions, a connection between genes and culture needs to be established. It is widely acknowledged that genes can influence culture, but the same does not go for the reverse. Instead, the consensus has generally been that “genes have culture on a leash. Culture can wander a bit, but if it threatens to get out of hand, its genetic master can bring it to heel.”

But Richerson and Boyd contend that “this is only half the story…. [H]eritable cultural variation responds to its own evolutionary dynamic, often leading to the evolution of cultural variants that would not be favored by selection acting on genes. The resulting cultural environments then can affect the evolutionary dynamics of alternative genes. Culture is on a leash, all right, but the dog on the end

---


is big, smart, and independent. On any given walk, it is hard to tell who is leading who.”

Richerson and Boyd maintain that genes and culture coevolve: “Genes, by themselves, can’t readily adapt to rapidly changing environments. Cultural variants, by themselves, can’t do anything without brains and bodies. Genes and culture are tightly coupled but subject to evolutionary forces that tug behavior in different directions.” Culturally acquired traits can affect the fitness of particular genotypes, and genotypes can affect the fitness of cultural traits. For example, deviations from culturally acquired social norms may be accompanied by punishment and a reduction in one’s chances to secure a mate. Likewise, genetic selection affects the kinds of traits that get passed on through culture. Genes are responsible for our psychology, and certain features of our psychology make the acquisition and transmission of cultural traits such as cooperation possible. Richerson and Boyd explain: “Each partner in the coevolutionary dance influences the evolutionary dynamics of the other. Genetically evolved psychological biases steer cultural evolution in genetic fitness-enhancing directions. Culturally evolved traits affect the relative fitness of different genotypes in many ways.”

26 Richerson and Boyd, Not by Genes Alone, 194.
27 Richerson and Boyd, Not by Genes Alone, 194.
28 Richerson and Boyd, Not by Genes Alone, 193.
29 Richerson and Boyd, Not by Genes Alone, 193.
In biology, “coevolution” refers to “systems in which two species are important parts of each other’s environments so that evolutionary changes in one species induce evolutionary modifications in the other.”\(^{30}\) Richerson and Boyd’s use of “coevolution” is different from the classical biological use of it. In the way that they use it, we are not dealing with two different species but a relation between genes and culture. Edward O. Wilson and Charles Lumsden were the first to apply the notion of coevolution to genes and culture.\(^{31}\) While this use of “coevolution” is not identical to its use in biology, there is nonetheless good reason to employ this terminology, since it brings to the foreground the interdependence between these two things that is captured in the genetic sense.

The “Baldwin effect” refers to a coevolutionary process whereby genetically-influenced behaviour can be shaped by culture. Terrence Deacon introduces it as follows:

Baldwin suggested that learning and behavioral flexibility can play a role in amplifying and biasing natural selection because these abilities enable individuals to modify the context of natural selection that affects their future kin. Behavioral flexibility enables organisms to move into niches that differ from those their ancestors occupied, with the consequence that succeeding generations will face a new set of selection pressures.\(^{32}\)

---

\(^{30}\) Richerson and Boyd, *Not by Genes Alone*, 192.


According to this process, culture can create an environment where genetic traits fare well. In some cases, learned behaviour permits an environmental change that subsequently results in the genetic selection of particular traits that are adaptive in this new culturally-created environment. This transition from culturally driven behaviours to genetic propensities itself is an adaptation, since genetic responses are generally more reliable. As Terence Deacon says, “developmentally ‘anticipating’ an important response by making its expression obligatory rather than contingent on external conditions provides a more efficient and less risky response than relying on an appropriate response being triggered by external signals.”

Deacon continues:

Of all the forms of adaptation, the flexibility to learn new behavioral responses during one’s lifetime can produce the most rapid and radical evolutionary consequences. Indeed, the ability to learn and thus inherit acquired behaviours may be one of the most powerful sources of evolutionary change. It provides an organism with access to a repertoire of potential adaptations, and so amplifies and extends the range of the behavioral predispositions that can be “sampled” by natural selection. A learned behavioral response can be genetically assimilated to become a behavioral predisposition by virtue of the costs that it imposes on the organism. There are costs in terms of learning time, costs for failing to learn or learning incorrectly, and costs for simply being inefficient. Individuals who, for any number of reasons, learn more quickly and reliably and implement this

33 Deacon, The Symbolic Species, 324.
behavior more efficiently will benefit in reproductive terms. One of the features of this process is that any predisposition that even remotely contributes to producing a more reliable and efficient response will be positively selected.34

Richerson and Boyd use lactose intolerance to illustrate this kind of gene-culture co-evolution.35 In most of the population, production of the enzyme needed to digest lactose shuts off after infancy. The advent of dairy farming, however, created an environment where the gene responsible for lactose digestion was favored. Cheese production, for example, permitted heavier reliance on milk products. This led to the migration of individuals to more mountainous areas—something that would not have been possible had they remained dependent on agriculture and required flat land. This new environment was one where the ability to digest lactose was adaptive, particularly in the vitamin-D deprived environment that individuals migrated to.36 And, “as that gene spread, it in turn may have changed the environment-shaping cultural practices, perhaps favoring more whole-milk consumption, or more serendipitously, giving rise to the evolution of ice cream.”37

34 Deacon, *The Symbolic Species*, 326.

35 Richerson and Boyd, *Not by Genes Alone*, 191-192. Although Richerson and Boyd do not explicitly identify it as such, this is an example of a Baldwin effect. Deacon describes the emergence of lactose tolerance in terms of Baldwinian evolution in *The Symbolic Species*, 323-324.

36 Another example is the prevalence of sickle cell anemia. The development of agriculture in parts of Africa created an environment conducive to malaria-carrying mosquitoes. Individuals with sickling of their cells were resistant to malaria and, thus sickle cells flourished. For a fuller discussion of this, see Deacon, 325-326.

37 Richerson and Boyd, *Not by Genes Alone*, 192.
A similar story can be told with respect to cooperation. Culture created an environment where prosocial dispositions were adaptive. As Richerson and Boyd say:

Cultural evolution created cooperative, symbolically marked groups. Such environments favored the evolution of a suite of new social instincts suited to life in such groups, including a psychology which “expects” life to be structured by moral norms and is designed to learn and internalize such norms; new emotions, such as shame and guilt, which increase the chance the norms are followed; and a psychology which “expects” the social world to be divided into symbolically marked groups.\(^{38}\)

And further:

As a result, people are endowed with two sets of innate predispositions, or “social instincts.” The first is a set of ancient instincts that we share with our primate ancestors. The ancient social instincts were shaped by the familiar evolutionary processes of kin selection and reciprocity, enabling humans to have a complex family life and frequently form strong bonds of friendship with others. The second is a set of “tribal” instincts that allow us to interact cooperatively with a larger, symbolically marked set of people, or tribe. The tribal social instincts result from the gene-culture coevolution of tribal-scale societies by the process described above. Consequently, humans are able to make common cause with a sizable, culturally defined set of distantly related individuals, a form of social organization that is absent in other primates.\(^{39}\)

---

\(^{38}\) Richerson and Boyd, *Not by Genes Alone*, 214.

\(^{39}\) Richerson and Boyd, *Not by Genes Alone*, 197.
Thus, individual-level selection works to generate a psychology that permits social behaviour. This selection is a result of the cultural environment shaped by selection at the group level. And, “if a process like cultural group selection worked over substantial periods of time, and especially if it recruited coercive strategies to effectively punish defectors from social contracts, it is reasonable to expect that human decision-making strategies should be adapted to life in a prosocial world.”

We thus get an evolutionary story of the origin of prosocial emotions. Richerson and Boyd say: “We think cultural evolutionary processes constructed a social environment that caused ordinary natural selection acting on genes to favor empathetic altruism, and a tendency to direct that altruism preferentially to fellow members of symbolically marked groups.” These emotions, in turn, help reinforce cooperative behaviour. As Bowles and Gintis suggest:

Some prosocial emotions, including shame, guilt, empathy, and sensitivity to social sanction, induce agents to undertake constructive social interactions; others, such as the desire to punish norm violators, reduce free riding when the prosocial emotions fail to induce sufficiently cooperative behavior in some fraction of members of the social group.

---


When cast in terms of selection on cultural rather than genetic variants, we can thus tell a plausible story about how groups will be organized in a way that is conducive to group selection. On Richerson and Boyd’s model, group selection on cultural variants is possible. Richerson and Boyd contend that cooperative norms will be ones that are more likely to survive, given the group-level advantages that they provide. And if so, cooperation can then be explained in terms of it being prescribed by norms that make groups do better when competing with other groups that employ different norms, as well as a set of emotional dispositions adapted to a cooperative cultural environment.
CHAPTER FIVE
Normative Social Contract Theory and
David Gauthier’s Contractarianism

Our concern thus far has been primarily with the descriptive enterprise of explaining the origins of cooperation in terms of evolutionary processes. I will, in this chapter, turn to the normative tradition, whose project is justificatory rather than explanatory. The normative tradition most concerned with cooperation—and consequently also the focus of this chapter—is the social contract tradition.

Approaches to Social Contract Theory

It will be useful to begin by situating the various and distinct approaches to social contract theory both in relation to one another and with respect to their central ambitions. We have seen that social contract theories derive moral obligations from an agreement based on reason. Individuals agree to the terms that will subsequently govern their social interaction with one another. We have also seen that social contract theories can be one of two types: (1) descriptive, and (2) normative. Both approaches are concerned with articulating the content of these terms of agreement, as well as with explaining and motivating compliance with them. The difference is that descriptive approaches are purely explanatory, whereas normative approaches are justificatory. Descriptive approaches explain the origin of current moral practices—
viz., how the current moral code of conduct originated and why individuals comply with it; normative approaches articulate the principles that *ought to* be in place and why individuals *should* comply with them.

My concern in this chapter will be with the normative branch of the social contract tradition, and more specifically with the contractarian arm of the normative tradition. More specifically still, my focus will be primarily on the problem of compliance, and on the options to overcome this problem that are available to the contractarian.

We have seen that the contractarian position is one of two branches within the normative social contract tradition, and lies in contrast to the contractualist view. Gauthier puts the distinction between the two branches in terms of the direction of justification: “On Scanlon’s [contractualist] view, the direction of justification is from the individual to others. The role of agreement is to gain the acceptance of others for one’s actions.”¹ The contractarian, on the other hand, “finds no basis for postulating a moral need for the justification of one’s actions to others. The role of agreement is to address each person’s demand that the constraints of society be justified to him, not a concern that he justify himself to his fellows.”²

This distinction can be traced to Rawls’s employment of the “rational” in the selection of the principles of justice in *A Theory of Justice*, and his subsequent shift in

---


emphasis to the “reasonable” in *Political Liberalism*. In Rawls’s view, human beings have two moral powers: a capacity for a sense of justice, and a capacity for a conception of the good. Rawls identifies the former with the “reasonable,” and the latter with the “rational.” According to Rawls, “persons are reasonable in one basic aspect when, among equals say, they are ready to propose principles and standards as fair terms of cooperation and to abide by them willingly, given the assurance that others will likewise do so.” The “reasonable,” for Rawls, “is an element of the idea of society as a system of fair cooperation and that its fair terms be reasonable for all to accept is part of its idea of reciprocity.” The “rational,” by contrast, “applies to how these ends and interests are adopted and affirmed, as well as to how they are given priority…what rational agents lack is the particular form of moral sensibility that underlies the desire to engage in fair cooperation as such, and to do so on terms that other as equals might reasonably be expected to endorse.”

Gauthier employs different terminology. According to him, the “maximizing” conception of rationality holds that, “the rational person still seeks the greatest satisfaction of her own interests.” According to the “universalistic” conception, by

---


contrast, “what makes it rational to satisfy an interest does not depend on whose interest it is.”\textsuperscript{9} Thus:

On the maximizing conception it is not interests in the self, that take oneself as object, but interests of the self, held by oneself as subject, that provide the basis for rational choice and action. On the universalistic conception it is not interests in anyone, that take any person as object, but interests of anyone, held by some person as subject, that provide the basis for rational choice and action. If I have a direct interest in your welfare, then on either conception I have reason to promote your welfare. But your interest in your welfare affords me such reason only given the universalistic conception.\textsuperscript{10}

There are thus two distinct conceptions of reason that can be endorsed that produce two distinct varieties of social contract theories. Vittorio Buffacci articulates the status of this conceptual landscape as follows:

Today, the debate between [social contract] theories of justice has degenerated into a shouting match: if you believe that we are moved by reasonableness and a sense of justice, then you will probably endorse something along the lines of a theory of justice as impartiality. Alternatively, if you believe that people are moved by rationality and self-regarding interests, then you will probably champion justice as mutual advantage. Advocates of justice as impartiality condemn Gauthier and his followers for misappropriating the language of morals to defend a position that has nothing to do with justice, while on their

\textsuperscript{9} Gauthier, \textit{Morals by Agreement}, 7.

\textsuperscript{10} Gauthier, \textit{Morals by Agreement}, 7.
part those who champion justice as mutual advantage frown on the likes of Rawls, Barry and Scanlon for making demands in the name of justice that are morally unwarrantable.\(^{11}\)

**Morality and Reason**

These two different conceptions of reason yield two very different answers to the question why one should be moral.\(^{12}\) David Gauthier has claimed that, “the reconciliation of morality and rationality is the central problem of modern moral philosophy.”\(^{13}\) One of the central obstacles for contract theories is to ensure stability of the contract—that is, to articulate a set of principles to which individuals will be willing and likely to comply. This problem is more acute for the contractarian branch, given its reliance on the maximizing conception of reason. If rationality is understood as the pursuit of one’s self-interest, and morality requires that we curb self-interested pursuits, then it is hard to see how rationality can be reconciled with morality. As Philippa Foot says, “justice seems rather to benefit others, and to work to the disadvantage of the just man himself.”\(^{14}\)

---


\(^{12}\) I remind the reader that I use the term “morality” here interchangeably with “cooperation,” and thus the question “Why be moral?” is equivalent to “Why cooperate?” This seems to be Gauthier’s main meaning. In “Bargaining Our Way Into Morality: A Do-It-Yourself Primer,” *Philosophic Exchange* 2, no. 5, (1979): 14-27, Gauthier distinguishes two parts of morality: “distributive justice” and “acquisitive justice.” The first concerns (as he puts it) “constraints on the modes of cooperation”; the second, “constraints on the baseline from which cooperation proceeds.” I will be concerned in this chapter only with the former.

The type of reason employed by contractualism, on the other hand, is straightforward to reconcile with moral constraints. But that conception of reason is difficult to motivate, and the background presuppositions about justice raise the question of how much work the notion of agreement is doing, and whether contractualism ought to properly be considered a social contract theory. By inserting moral arguments into the choice schema, it becomes less clear that the principles that are arrived at are the result of agreement rather than arise primarily from the prior set of values that convince the parties to accept that agreement.15

Gauthier articulates this difficulty very nicely: either one’s conception of reason is hard to square with morality or it is hard to defend.

Morality…is traditionally understood to involve an impartial constraint on the pursuit of individual interest. The justification for such a constraint poses no problem to the proponents of universalistic rationality. The rational requirement that all interests be satisfied to the fullest extent possible directly constrains each person in the pursuit of her own interests.16


15 Two central objections that face Scanlon’s contractualist view that emerge from his reliance on the reasonable rather than the rational, for example, are redundancy and circularity. Both of these call into question the utility of the social contract metaphor. Questions about what is fair are either determined prior to being inputted into the contract machinery (and thus the contract is redundant), or require justification that is dependent on the contract itself (and thus the contract procedure is circular). On Scanlon and redundancy, see further: Brad Hooker, “Contractualism, Spare Wheel and Aggregation,” in Scanlon and Contractualism, ed. Matt Matravers (London: Frank Cass, 2003), 53-76; Philip Pettit, “Doing Unto Others,” Times Literary Supplement (June 25, 1999): 7-8; Philip Stratton-Lake, “Scanlon’s Contractualism and the Redundancy Objection,” Analysis 63, no. 1 (January 2003): 70-76. On Scanlon and circularity, see further: Gerald Dworkin, “Contractualism and the Normativity of Principles,” Ethics 112, no. 3 (April 2002): 471-482; Aaron James, “Rights and Circularity in Scanlon’s Contractualism,” Journal of Moral Philosophy 1, no. 3 (2004): 367-374.
He goes on to say that:

The main task of our moral theory—the generation of moral constraints as rational—is thus easily accomplished by proponents of the universalistic conception of practical reason. For them the relation between reason and morals is clear. Their task is to defend their conception of rationality, since the maximizing and universalistic conceptions do not rest on equal footings. The maximizing conception possesses the virtue, among conceptions, of weakness. Any consideration affording one a reason for acting on the maximizing conception, also affords one such a reason on the universalistic conception. But the converse does not hold. On the universalistic conception all persons have in effect the same basis for rational choice—the interests of all—and this assumption, of the impersonality or impartiality of reason, demands defence.\footnote{Gauthier, \textit{Morals by Agreement}, 7.}

Gauthier writes further that:

\[F\]ew persons would embrace the universalistic conception of practical reason did they not think it necessary to the defence of any form of rational morality. Hence the most effective rebuttal of their position may be, not to seek to undermine their elaborate and ingenious arguments, but to construct an alternative account of a rational morality grounded in the weaker assumption of the theory of rational choice.\footnote{Gauthier, \textit{Morals by Agreement}, 7-8.}
I will follow Gauthier in this, and defend the rational conception of reason (and thus contractarianism). And thus if this defense is successful there will be no need to look at contractualism.

Central to moral and political philosophy (and particularly emphasized in social contract theory) are questions of the content of cooperative norms as well as the motivation to comply with those norms. We have seen that the contractarian approach, embodied in the work of Hobbes, grounds these projects in reason. Probably the most well-known and well-developed contemporary attempt to revive Hobbesian contractarianism is that of David Gauthier. Gauthier articulates his aim as follows:

We shall develop a theory of morals as part of the theory of rational choice. We shall argue that the rational principles for making choices, or decisions, among possible actions, include some that constrain the actor pursuing his own interest in an impartial way. These we identify as moral principles.\(^{19}\)

Gauthier’s project can be set out in two major stages.\(^{20}\) In the first stage, he sets out the terms under which rational individuals will agree to cooperate with one another. Gauthier argues that the principle of justice that rational individuals will select is the


\(^{20}\) I will ignore here Gauthier’s third aim of establishing what he calls the “Lockean Proviso,” which sets the appropriate baseline for cooperation. On this, see further Gauthier, *Morals by Agreement*, 190-232.
Principle of Minimax Relative Concession (MRC).\textsuperscript{21} In the second stage, Gauthier seeks to show that it is rational to comply with the arrangement determined in the first stage even when violating it might be advantageous. Gauthier locates the rationality of compliance with agreements in the adoption of the disposition he refers to as “constrained maximization.”\textsuperscript{22}

My concern is primarily with the second stage of Gauthier’s argument—that is, with his argument for the rationality of constrained maximization. The following brief sketch of the first stage will suffice for my purposes.

**The Bargaining Problem and Minimax Relative Concession**

One of the central aims of social contract theory is to show that rational agents (however they are idealized) will endorse the principles of justice put forth by their theories. The starting point is to characterize the contracting agents, and then show that these agents will endorse a particular set of principles. These become the principles of justice, and are rendered legitimate in virtue of the fact that they are the principles that would be selected by rational agents idealized under certain conditions.

Gauthier sets these terms as an outcome of a hypothetical bargaining situation. When individuals engage in mutually advantageous cooperative arrangements, a surplus is generated from that interaction. This is the amount of utility that is generated above and beyond what each individual brings to the initial imagined

\textsuperscript{21} Gauthier, *Morals by Agreement*, 145. For a more detailed discussion of this, see 113-157.

\textsuperscript{22} See Gauthier, *Morals by Agreement*, 157-189.
bargaining position. What each person takes away from the bargaining table must be an improvement over what she began with. Each person then advances a claim. Rational individuals, as utility-maximizers, will begin by demanding as much as possible and, thus, their claims will inevitably conflict. They will consequently have to withdraw some of their claim and make a concession. The absolute magnitude of a concession is the difference between an individual’s initial claim and how much she actually gets from the bargaining solution. Gauthier defines the relative concession as “the proportion its absolute magnitude bears to the absolute magnitude of complete concession.”

Gauthier contends, in Morals by Agreement, that rational individuals will settle on the solution that yields the minimum maximum relative concession for each player. This is Gauthier’s Minimax Relative Concession principle (MRC), which he formulates as:

In any co-operative interaction, the rational joint strategy is determined by a bargain among the co-operators in which each advances his maximal claim and then offers a concession no greater in relative magnitude than the minimax concession.

My aim here is not to examine the adequacy of Gauthier’s solution. It is, however, worthwhile to make two points. First, Gauthier concedes in a later paper

---

23 Gauthier, Morals by Agreement, 136.

24 Gauthier, Morals by Agreement, 145.

25 There have been several lines of objection to Gauthier’s derivation of MRC. Gauthier outlines these as follows: (1) Rational bargaining does not yield MRC, but rather the Nash bargaining solution; (2)
that his bargaining solution cannot stand as it was presented in *Morals by Agreement*.

He says:

The argument in Chapter V of *Morals by Agreement* cannot stand in its present form. At most it may have heuristic value in presenting the idea of minimax relative concession as uniting rationality and morality. But the real work of defending MRC as a bargaining outcome, if as I still believe it is defensible, requires a different argument. It requires arguing that in the circumstances of the social contract, MRC coincides with the Nash bargaining solution.26

Second, the central objection addressed by Gauthier in the above, namely that rational bargaining does not result in MRC—that is, rational individuals in a bargaining situation would not settle on MRC as the solution to their bargaining problem—illustrates a more general problem (that will be familiar from our discussion of equilibrium selection) for the enterprise of deriving a unique set of principles of justice from reason. This is that there is some indeterminacy with respect to what solution rational bargaining gives rise to. As Peter Vanderschraaf says:


Nash (1953) argued that any bargaining problem can be analyzed either from the perspective of an arbiter, who employs certain axioms to determine a just division, or from the perspective of the agents themselves, who try to resolve the problem via some bargaining process. Braithwaite, and Gauthier (1986) after him, took the axiomatic approach. They defend a solution proposed by Raiffa (1953) and Kalai-Smorodinsky (1975) which divides the resource so that if the payoffs for each is 0 at the (H, H) outcome, then the ratio of payoffs an agent receives to his payoffs if he gets all of the resource is the same for each agent. But as Luce and Raiffa admit (1957), the arguments for the Raiffa-Kalai-Smorodinsky solution are by no means decisive. Nash proposed a different axiomatic solution for the two-agent bargaining problem, according to which the resource is divided so as to maximize the product of the two agents’ payoffs given that they both receive 0 if both follow H. A utilitarian solution would divide the resources so as to maximize the sum of the payoffs. And there are still other axiomatic solutions defended in the literature. There is no generally accepted solution to the general two-agent bargaining problem, let alone the general bargaining problem for two or more agents.  

Thus, the fact that there is no unique solution (or no single widely endorsed solution) to the bargaining problem reveals a deficiency in rational choice theory, and casts doubt on the adequacy of accounts that rely on rationality to derive their principles of justice. However, more pressing than the above is the compliance problem, to which I

---

will now turn. For if it cannot be shown that it is rational to comply with the terms of cooperation, that would be fatal to the enterprise, no matter what solution we arrive at in the bargaining problem.

**The Compliance Problem**

Even if Gauthier could show that rational individuals would settle on MRC as the appropriate solution to the bargaining problem (or if some other suitable solution could be found), there remains the problem of compliance. Recall that in Prisoner’s Dilemma cases mutual cooperation is preferable to mutual defection. But this preferred outcome of mutual cooperation to mutual defection is not in equilibrium, as each player’s most preferred outcome is defection while the other player complies. Thus, Gauthier must show that it is not only rational to *make* agreements in accordance with the principle of MRC, he must also show that it is rational to *comply* with agreements made in accordance with that principle.

The standard contractarian solution to the problem of reconciling morality with rationality suggests that the incompatibility is only apparent. Human beings are not entirely self-sufficient and rely for their well-being on social interaction. Peaceful social interaction is established through moral norms. Thus, morality serves one’s interest by enabling one to partake in social interaction. The rules of morality do require one to curb self-interested pursuits but, as the argument goes, this restraint is necessary to further one’s overall interest, so long as others are willing to do the same. As Kurt Baier puts it, “moralities are systems of principles whose acceptance
by everyone as overruling the dictates of self-interest is in the interest of everyone alike, though following the rules of morality is not of course identical with following self-interest.”

On the contractarian view, the justification for the constraints characteristic of moral obligations then lies in the role that they play in facilitating the pursuit of individual interest.

Hobbes thought that he could derive cooperative behaviour from the minimal assumptions that human beings were both rational and self-interested. According to Hobbes, human action is selfishly motivated, and reason shows us the means to such self-interested ends. It is in virtue of these qualities that Hobbes argues that human beings will be able to make agreements constitutive of morality. Sufficiently rational and self-interested individuals will recognize that the benefits of cooperative behaviour far exceed those of non-cooperation. His argument goes like this. Left on their own, individuals will pursue their own interests. Given this, and that the means by which individuals can secure their own survival are limited, it is inevitable that they will come into conflict with one another. Person A will want to take what person B possesses, and person B will want to take what person A has. Recognizing the possibility of attack from B, A will make a pre-emptive strike against B, and B will reason similarly and behave analogously towards A. Individuals will thus be in a state of war, “where every man is enemy to every man,” and where there can be no assurance of survival. Thus, everyone can do better if they make agreements of non-


aggression. Individuals agree to cooperate with one another (i.e., agree to not take the other’s possessions, and so on) so long as others are willing to do the same. Given Hobbes’s conception of the rationally selfinterested nature of human beings and the rough equality that exists between them, the absence of cooperation will inevitably lead to a state where no one’s interests are furthered nor even where survival can be secured. Thus, rational persons will make cooperative agreements to avoid the alternative and much worse life in the state of nature.

Those who think cooperation can be reconciled with rationality will point to the fact that mutual cooperation will yield a higher utility than will mutual defection. But proponents of a strictly maximizing conception of rationality will contend that in the Prisoner’s Dilemma to cooperate is a dominated strategy (that is, no matter what one’s opponent does, defection is always the best reply in terms of utility-maximization) and is thus positively irrational.

Hobbes’s Foole takes this line, and asks why one cannot violate agreements in cases where doing so is advantageous.

The foole hath said in his heart, there is no such thing as Justice; and sometimes also with his tongue; seriously alleging, that every man’s conservation, and contentment, being committed to his own care, there could be no reason, why every man might not do what he thought conduced thereunto: and therefore also to make, or not make; keep, or not keep covenants, was not against reason, when it conduced to one’s benefit. 30

The Foole’s objection points to a structural problem in the Prisoner’s Dilemma: it will always be to one’s benefit to violate one’s agreement. And insofar as one is rational to the extent that one pursues one’s benefit, the rational course of action will always be to defect.

In reply to the Foole, Hobbes argued that violations of agreements are liable to be detected and punished, and that the consequences of being caught—that is, being excluded from civil society—are so grave that defection is never a good gamble.31 This reply in effect changes the payoff structure of the Prisoner’s Dilemma such that unilateral defection carries with it grave consequences rather than rich rewards. But in order for this reply to be successful, the highly unlikely state of affairs must obtain where, for every single potential defection, the risks and possible costs associated with defection outweigh any possible gains. Gauthier says, “Hobbes needs to say that it is rational to perform one’s covenant even when performance is not directly to one’s benefit, provided that it is to one’s benefit to be disposed to perform. But this he never says. And as long as the Foole is allowed to relate reason directly to benefit in performance, rather than to benefit in the disposition to perform, he can escape refutation.”32 Hobbes’s ultimate solution to the problem of non-compliance is a political one: to have a sovereign with sufficient power of surveillance and authority to punish so as to make non-compliance counterproductive. This, however, can be


32 Gauthier, Morals by Agreement, 162.
costly and inefficient, and it would be desirable if compliance could be achieved by more efficient non-coercive means.

Gauthier presents us with such a means. He locates the rationality of compliance with agreements in the adoption of the disposition he refers to as “constrained maximization.”

Constrained maximizers conditionally dispose themselves to cooperation. This disposition to cooperation distinguishes the constrained maximizer from what Gauthier refers to as a “straightforward maximizer,” who “seeks to maximize his utility given the strategies of those with whom he interacts.” Straightforward maximizers (like the Foole) are rational utility-maximizers; they will defect when it is advantageous for them to do so. The constrained maximizer, by contrast, will cooperate when he expects others to cooperate and defect only when he anticipates that others will do the same.

The underlying presupposition of constrained maximization is the rationality of constraining utility-maximization in order to pursue a mutually optimal strategy. The constrained maximizer disposes himself, essentially, to forgo token opportunities to make big gains through defection in order to obtain the benefits of mutual cooperation. Gauthier must then show that doing so is rational. This requires showing that the gains through mutual cooperation outweigh gains through defection, a conclusion that is not obvious given the structure of interactions between the two types of maximizers.

---

33 Gauthier, Morals by Agreement, 167.

34 Gauthier, Morals by Agreement, 167.
In a society composed of a mix of straightforward and constrained maximizers, individuals will defect in all but two cases. First, two constrained maximizers interact and are able to identify each other as constrained maximizers. In such a case, both will cooperate. Second, a constrained maximizer mistakes a straightforward maximizer for a constrained maximizer. In such a case, the constrained maximizer will adhere to the agreement, and the straightforward maximizer will defect. In Prisoner’s Dilemmas, the most utility-maximizing strategy is unilateral defection. If the straightforward maximizer is able to fool the constrained maximizer into thinking that he is a constrained maximizer, then he will be able to gain at the expense of the constrained maximizer. This will yield the best outcome for the straightforward maximizer and the worst for the constrained maximizer. The second best outcome for both is mutual adherence, and the third best for both is mutual defection.

When neither party cooperates, both the straightforward and the constrained maximizer will receive the third best payoff. In an encounter between a straightforward maximizer and a constrained maximizer, if the constrained maximizer adheres to the agreement while the straightforward maximizer defects, the straightforward maximizer will receive a payoff greater than he would had he adhered. The straightforward maximizer is thus able to reap advantages unavailable to the constrained maximizer through unilateral defection. Although the constrained maximizer has available to him opportunities for gain through mutual adherence, mutual adherence yields a utility less than does unilateral defection.
In order to avoid the impending conclusion that rational individuals ought always to defect, Gauthier must rely on the assumption that straightforward maximizers cannot pass as constrained maximizers. For, given that straightforward maximizers will do better than constrained maximizers as long as they are given the same opportunities as constrained maximizers, and whether one is able to partake in agreements with others depends on whether it appears that one can be trusted, so long as one is able to maintain the illusion of being a constrained maximizer, one will be able to partake in agreements with others while reaping the benefits of defection. Straightforward maximizers thus will do better than constrained maximizers, which will prevent Gauthier from claiming that constrained maximization is rational.

Gauthier recognizes how crucial the detection of dispositions and intentions is. If people were (to use Gauthier’s terminology) transparent (i.e., their characters were always accurately detectable by others), then constrained maximization would be the rational strategy. For in this case constrained maximizers would be able to identify straightforward maximizers, and consequently exclude them from agreements. If people were, on the other hand, what Gauthier describes as opaque (i.e., their characters remained hidden to others), then straightforward maximizers would do better. For straightforward maximizers would then be able to continue to make agreements with others and gain through successful exploitation. Gauthier argues that neither of these is realistic, and claims that people are translucent, according to which “persons are neither transparent nor opaque, so that their

\[35\] Gauthier defines this and the following related terms “opaque” and “translucent” in Morals by Agreement, 173-174.
disposition to co-operate or not may be ascertained by others, not with certainty, but as more than mere guesswork.” In other words, if our characters are translucent, it is assumed that one has a better chance of correctly identifying another’s character than they do at random.

If so, and if one who is suspected to be a straightforward maximizer will be excluded from cooperative agreements, then the straightforward maximizer will not reap all projected benefits of both cooperation and defection. The straightforward maximizer will forgo many benefits of cooperation and reap only those gains that he can through successful exploitation. As his untrustworthy character becomes more widely known, opportunities for exploitation will diminish. And since constrained maximization affords one the possibility of gaining through mutually-beneficial cooperative interactions—opportunities unavailable to the straightforward maximizer—it becomes plausible to suggest that constrained maximization will yield a higher utility to those who so dispose themselves than will straightforward maximization. Gauthier says:

Both CMs [constrained maximizers] and SMs [straightforward maximizers] must expect to benefit from increasing their ability to detect the dispositions of others. But if both endeavour to maximize their abilities (or the expected utility, net of costs, of so doing), then CMs may expect to improve their position in relation to SMs. For the benefits gained by SMs, by being better able to detect their potential victims, must be on the whole offset by those

---

losses they suffer as the CMs become better able to detect them as potential exploiters. On the other hand, although the CMs may not enjoy any net gain in their interactions with SMs, the benefits they gain by being better able to detect other CMs as potential co-operators are not offset by corresponding losses, but rather increased as other CMs become better able to detect them in return.\textsuperscript{37}

Gauthier thus concludes that rational persons will become constrained maximizers.

It is important to note that Gauthier does not contend that constrained maximization will \textit{always} be rational:

If persons are translucent, then constrained maximizers (CMs) will sometimes fail to recognize each other, and will then interact non-co-operatively even if co-operation would have been mutually beneficial. CMs will sometimes fail to identify straightforward maximizers (SMs) and will then act co-operatively; if the SMs correctly identify the CMs they will be able to take advantage of them. Translucent CMs must expect to do less well in interaction than would transparent CMs; translucent SMs must expect to do better than would transparent SMs. Although it would be rational to choose to be a CM were one transparent, it need not be rational if one is only translucent.\textsuperscript{38}

\textsuperscript{37} Gauthier, \textit{Morals by Agreement}, 181.

\textsuperscript{38} Gauthier, \textit{Morals by Agreement}, 174. Based on calculations of expected utility in varying situations, Gauthier draws the following two conclusions:

First, it is rational to dispose oneself to constrained maximization only if the ratio of $p$ to $q$, i.e., the ratio between the probability that an interaction involving CMs will result in co-operation and the probability that an interaction involving CMs and SMs will involve exploitation and defection, is greater than the ratio between the gain from defection and the gain through co-operation… Second, as the proportion of CMs in the population increases (so that the value of $r$ [the probability that a randomly selected member of the population will be a CM] increases) the value of the ratio of $p$ to $q$ that is required for it to be rational to dispose oneself to constrained maximization decreases (176).
The rationality of constrained maximization will depend on the payoffs associated both with successful exploitation and with cooperation, as well as on the degree to which individuals are able to detect the dispositions of others. As Gauthier says:

If we fall into a society—or rather into a state of nature—of straightforward maximizers, then constrained maximization, which disposes us to justice, will indeed be of no use to us, and we must then consult only the direct dictates of our own utilities. In a world of Fooles, it would not pay to be a constrained maximizer, and to comply with one’s agreements. In such circumstances it would not be rational to be moral.39

Thus, sometimes the disposition to constrained maximization will yield a lower utility than will the disposition to straightforward maximization, and so it will not always be rational to dispose oneself to constrained maximization. But so long as people are translucent enough, and the payoffs of cooperation are appropriately balanced to those of unilateral defection, then the disposition to constrained maximization is rational. In Gauthier’s words:

[I]f we find ourselves in the company of reasonably just persons, then we too have reason to dispose ourselves to justice. A community in which most individuals are disposed to comply with fair and optimal agreements and practices, and so to base their actions on joint co-operative strategies, will be self-sustaining. And such a world offers benefits to all which the Fooles can never enjoy….Only the person truly disposed to honesty and justice may

expect fully to realize their benefits, for only such a person may rationally be admitted to those mutually beneficial arrangements—whether actual agreements or implicitly agreed practices—that rest on honesty and justice, on voluntary compliance….The Foole and the sensible knave, seeing the benefits to be gained from the exceptions, from the advantageous breaches in honesty and compliance, but not seeing beyond these benefits, do not acquire the disposition. Among knaves they are indeed held for sensible, but among us, if we be not corrupted by their smooth words, they are only fools.

Robert Sugden says that, “in voluntarily imposing constraints on herself—in being willing to reject the action that maximizes her expected utility—the CM is acting in a recognizably moral way. But the disposition to act in this way is utility-maximizing and thus rational.” And as Gauthier puts it:

The idea is that in real interaction it is reasonable to accept co-operative arrangements that fall short of the ideal of full rationality and fairness, provided they do not fall too far short. At some point, of course, one decides to ignore a joint strategy, even if acting on it would afford one an expected utility greater than one would expect were everyone to employ an individual strategy, because one hopes thereby to obtain agreement on, or acquiescence in, another joint strategy which in being fairer is also more favourable to oneself. At precisely what point one decides this we make no attempt to say. We simply defend a conception of constrained maximization that does not require that all acceptable joint strategies be ideal….Constrained

---

40 Gauthier, Morals by Agreement, 182.

maximization thus links the idea of morals by agreement to actual moral practice.\textsuperscript{42}

Thus, to the extent that forming a disposition to constrained maximization resembles forming a disposition to cooperate, if Gauthier’s defense of constrained maximization is successful, then he will have provided an account that reconciles cooperative behaviour with rationality. Let us now turn to evaluate the success of his account.

**Rationality and Constrained Maximization**

The success of Gauthier’s reconciliation project depends primarily on the answers to two questions. The first is whether constrained maximization is rational. The second is whether acting on that disposition is rational. I will take these in turn.

**Is it Rational to Become a Constrained Maximizer?**

The rationality of adopting a disposition depends on how well it does relative to other dispositions. In order to establish that it is rational to dispose oneself to constrained maximization, Gauthier must show that individuals so disposed will do better than those disposed to straightforward maximization. Gauthier must also show that constrained maximizers do better than those maintaining other dispositions.

David Copp suggests the disposition of “reserved maximization” as an alternative to constrained maximization.\textsuperscript{43} A reserved maximizer will act much like

\textsuperscript{42} Gauthier, *Morals by Agreement*, 168, my emphasis.
the constrained maximizer in most circumstances, but will violate an agreement in certain cases where the constrained maximizer will adhere, such as when he stands to gain substantially and the likelihood of detection is significantly low. Since one will almost always act like a constrained maximizer, one will gain the benefits of cooperation. But since one will not absolutely close the door on a “big score,” one will also have some of the benefits of a straightforward maximizer.

Constrained maximizers are able to reap rewards of cooperation. So are straightforward and reserved maximizers, as long as they are not detected. Straightforward and reserved maximizers are also able to reap rewards of defection, again so long as they are not detected. So whether constrained maximization does better than alternative dispositions ultimately hinges on whether or not straightforward or reserved maximizers are detected. The rationality of becoming a constrained maximizer thus depends on whether humans are (or can become) translucent.

The plausibility of translucency seems to be largely an empirical matter, and there is an emerging literature on cheater detection abilities in humans that may help put to rest concerns about translucency. Gauthier could point to physiological

---


reactions that accompany deceit—rapid heartbeat, flushing, aversion of eye contact, and so on, to support the view of the detectability of dispositions. He could also say that we can determine the dispositions of others in more impersonal ways by examining the history of their behaviour. Edward McClennen writes:

As it turns out, the iterated game framework provides a setting in which the epistemological problem of assurance can be resolved. If interaction is sufficiently ongoing, then for many kinds of encounters, a given individual can have the requisite assurance regarding the disposition of other participants. The history of past encounters between participants typically will provide the needed information. It is plausible to suppose, moreover, that for many such contexts, the requisite information will be securable from anecdotal sources—that is, it will be unnecessary to resort to formal mechanisms for the compiling and transmission of this information. At the “street level,” each typically will be able to consult personal experience and informally shared information with friends and family members to determine whether or not the level of voluntary cooperation in more impersonal, “public” settings has been great enough to warrant voluntary compliance on one’s own part.  

Critics will say otherwise, and may point to instances where individuals successfully lie to or cheat their fellow men. The issue is yet to be settled, but as it stands, there is no real knockdown argument against translucency, and I thus conclude that

---

Gauthier’s argument for constrained maximization remains undefeated on the ground of the implausibility of translucency.

Furthermore, there is reason to believe that the evolutionary story of the origins of human cooperation can help to bolster the view that humans have evolved as translucent creatures. If the story I have endorsed about group selection on cooperative cultural variants and individual selection of prosocial dispositions is defensible, then it would be plausible to suggest that among those prosocial dispositions would have also evolved an ability to detect the character of others and signal one’s own trustworthy character to others.

Rational Dispositions and Rational Actions

It has been suggested that even if it can be established that those who develop a disposition to constrained maximization do better than those who do not, that does not entail that it is also rational to act on that disposition. In other words, the success of Gauthier’s project rests not only on whether it is rational to acquire the disposition to constrained maximization, it must also be established that it is rational to act on that disposition. Critics of Gauthier might grant that it is rational to form a disposition to constrained maximization but deny that it is rational to carry through with that

---

46 See, for example, Gregory Kavka, “Review: Morals by Agreement,” *Mind* 96, no. 381 (January 1987): 117-121 and Derek Parfit, “Bombs and Coconuts, or Rational Irrationality,” in *Practical Rationality and Preference*, 81-97. Both Kavka and Parfit contend that to form a disposition to constrained maximization is to form a disposition to perform irrational acts. On their view, Gauthier cannot claim that it is rational to carry through with actions recommended by constrained maximization and, thus, cannot successfully carry out his reconciliation project. For the reasons I will outline in this chapter, I contend that Gauthier does succeed in defending the rationality of actions recommended by constrained maximization.
disposition. Consider Gauthier’s example of harvesting one’s crops, borrowed from Hume.\textsuperscript{47} Persons A and B have crops to harvest. They can do so alone or they can agree to help one another. Helping one another will involve one person helping the other at T1, and the other returning the help at T2. In order for person A to receive assistance from person B in harvesting her own crops, she will have to assure B that, after B helps her at T1, she will provide B with assistance at T2. Suppose that B helps A at T1. Is it now, at T2, rational for A to help B?

Critics of Gauthier will contend that it is not. Their argument might go like this. A’s act of assuring B constitutes A forming a disposition to help B. Assuming translucency, A’s forming of the disposition causes B to help A. A then gains from the disposition—a gain she would not have obtained had she not formed the disposition. It is thus rational for A to form the disposition to help B. But now comes T2. Is it now rational for A to actually do what she disposed herself to do? Helping B at T2 imposes a cost on A. At T2 A has already received B’s help and so (let us suppose) has nothing to further gain from helping B at T2. Thus, A has no reason to help B at T2. How can we say that helping B is now rational?

To relate this example to constrained maximization, while it might be rational to dispose oneself to constrained maximization at T1, insofar as doing so secures a cooperative gain at T2, it is not clear that at T2 carrying through with one’s part of the agreement is rational. At T2, having already secured the cooperative outcome, the

individual disposed to constrained maximization has nothing further to gain from that disposition. At T2, it is utility-maximizing for an individual to act as a straightforward maximizer. Thus it would seem rational to, at T2, abandon actions recommended by constrained maximization.

Gauthier resists this conclusion and argues that the rationality of forming the disposition to constrained maximization does indeed entail the rationality of actions recommended by that disposition. According to him, there is a crucial relationship between an agent’s assurance that she will carry through with her agreement, her intention to do so, and her success in securing the cooperative outcome. Specifically, Gauthier contends that in order for A to receive the cooperative benefit from B at T2, A must provide B at T1 with a sincere assurance that she will carry through with the agreement. In order to provide a sincere assurance at T1, A must have the intention that at T2 she will carry through with the agreement. In order to have the intention to cooperate at T2, A must believe that it is rational for her to do so. According to Gauthier, A cannot, without inconsistency, provide B with sincere assurance that she will cooperate at T2 if she knows that at T2 it will no longer be to her benefit (and will thus be irrational) to cooperate.

---

48 Gauthier presents his argument to this effect in “Assure and Threaten.”

49 In “Assure and Threaten,” Gauthier says:

To be sure, my life would go better were I to give you the sincere assurance and then not honor it. But although I can do these things—I can be perfectly sincere in assuring you that I shall return your help and yet not return it—I cannot choose to do them taken together. I cannot simultaneously choose both to give you a sincere assurance to return your help and not to honor my assurance when the time comes. Furthermore, if I choose to give you a sincere assurance, then in so choosing I must intend to honor it and believe that I shall honor it (695).
But if A evaluates the rationality of her actions at the level of action, she must concede that it is not rational to cooperate at T2, since at T2 cooperating is not the utility-maximizing option. If A evaluates her action this way, then she cannot make a sincere assurance at T1 to cooperate at T2, since she knows that at T2 cooperating is not the utility-maximizing (and thus rational) option. Without the sincere assurance that A will cooperate, B will not either. Both A and B will end up with the non-cooperative outcome, and each will harvest her crops alone.

To avoid this outcome, Gauthier recommends the adoption of a different perspective from which to evaluate the rationality of action. His concern shifts to the rationality of dispositions rather than actions. According to Gauthier, “intentional structures create problems for the orthodox account of deliberation, which insists that rational actions are those that directly promote the agent’s aim, taking as illustrative the aim that one’s life go as well as possible.” He says, “if my aim is that my life go as well as possible, then I should not take all of my reasons for acting directly from that aim, considering only which action will have best consequences for my life. For if I always deliberate in this way, then my life will not go best for me.”

Gauthier’s revised account of rational deliberation permits A to cooperate at T2 so long as making the assurance at T1 to cooperate at T2 yields a greater utility than doing otherwise. And, “since the direct link between rational deliberation and particular outcomes has been severed, an action may be rational even though at the

---


time of performance it is not, and is not believed to be, part of a life that goes best for
the agent.\(^{52}\) Thus, it is rational for A to cooperate at T2.

Thus, on the above account of rational deliberation, good reasons are those
that lead to optimality. Gauthier says:

In defending constrained maximization we have implicitly reinterpreted the
utility-maximizing conception of practical rationality. The received
interpretation, commonly accepted by economists and elaborated in Bayesian
decision theory and the Von Neumann-Morgenstern theory of games,
identifies rationality with utility maximization at the level of particular
choices. A choice is rational if and only if it maximizes the actor’s expected
utility. We identify rationality with utility-maximization at the level of
dispositions to choose.\(^{53}\)

This reconceptualization of reason permits Gauthier to say that an action
recommended by constrained maximization is rational, even if there is an alternative
act that will yield a higher utility so long as the disposition to constrained
maximization yields greater utility than does the disposition recommending the
alternative action (in our case, the disposition to straightforward maximization).\(^{54}\)

\(^{52}\) Gauthier, “Assure and Threaten,” 701.


\(^{54}\) David Copp says:

Gauthier’s ingenious solution involves moving from the appraisal of one’s choices to the
appraisal of one’s dispositions to choose. This reorientation of rational assessment,
coordinated with a key amendment to the theory that rational choice consists in
straightforward expected utility maximization, is meant to yield the desired result. He argues
that one can maximize one’s expected utility in choosing to form a specific conditional
Actions that, if taken as individual tokens are non-utility-maximizing are rational, so long as they are recommended by a disposition that it is rational to have. This account resolves the problem of compliance and ensures a superior outcome to one obtainable from a maximizing conception.

One might ask whether Gauthier is entitled to make this kind of reconceptualization of rationality. McClennen\textsuperscript{55} thinks so. On his view, while the cooperative outcome is a dominated and non-equilibrium point in the Prisoner’s Dilemma, that fact alone should not exclude that outcome as a resolution to the problem. According to him, convergence can in fact occur at loci other than equilibrium points, and he thinks that Pareto-Optimality\textsuperscript{56} is one such locus. He considers games of pure coordination, i.e., games where there is no conflict of interest between players and the goal is merely to coordinate their strategies. Of these games, he says:

\begin{quote}
[T]he appropriate concern for a rational players \textit{sic.}….is to coordinate strategies so that the outcome will satisfy the Pareto condition. It is true, of course, that outcomes satisfying the Pareto condition satisfy the equilibrium disposition to conform with certain constraints on maximizing activity, viz., constrained maximization. Because this is so, constrained maximization is a rational disposition, and, Gauthier claims, any action that “expresses” this disposition also counts as rational, even if it involves, as any such action might, the failure to maximize one’s expected utility.
\end{quote}

\hspace{1cm}“Contractarianism and Moral Skepticism,” in \textit{Contractarianism and Rational Choice}, 200.

\textsuperscript{55} McClennen, “The Strategy of Cooperation,” 189-208.

\textsuperscript{56} Gauthier defines this as follows: “An (expected) outcome is optimal (or, more fully, Pareto-optimal) if and only if there is no possible outcome affording some person a greater utility and no person a lesser utility. Alternatively, an outcome is optimal if and only if any other outcome affording some person a greater utility also affords some other person a lesser utility.” \textit{Morals by Agreement}, 76.
condition. But from the perspective of the strategic problem that players face in a game of pure coordination, this additional property is purely accidental. That is, the equilibrium concept adds nothing that illuminates the nature of the deliberation that persons face in such games. In this context, it does no work.  

He continues:

To settle…on a suboptimal outcome…is to settle for less than one could have, under conditions in which the other player is also settling for less than he or she could have. Both, that is, forgo an opportunity, which both players would prefer, and in the absence of a special reason why they should collectively forgo such a mutually preferred outcome, passing up such an opportunity would be irrational.

On McClennen’s view, since in coordination games it is not clear that the equilibrium concept itself plays much of a role in the determination of strategic action, and since instead much of the work is (and, according to him, ought to be from the point of view of rationality) done by Pareto-considerations, there is reason to suppose that the same might go for cooperation problems. And thus:

[I]deally rational persons who have common knowledge of their rationality and the game they are playing will be disposed to cooperate to ensure a Pareto-efficient and Pareto-optimal outcome. What this means is that rational agents who know each other to be such will not confine their attention to


equilibrium combinations of choices. Specifically, if parametric or equilibrium reasoning leads to a suboptimal outcome, and there is an unambiguous combination of coordinated choices that is Pareto-efficient relative to the equilibrium outcome, then they will coordinate on it….Within the context of games in which there is an outcome that uniquely satisfies the Pareto conditions, and under conditions of common knowledge, Pareto considerations will prevail. In particular, then, under conditions of common knowledge, rational players will cooperate in one-shot sequential prisoner’s dilemma games.  

The door is thus opened to an account like Gauthier’s where reason is reconceptualized to guarantee the rationality of actions recommended by rational dispositions.  

This is not as unusual a move as it may seem. There is a parallel discussion in the literature on consequentialism. Consider, for example, the “Paradox of Hedonism.” An unadulterated version of hedonism would require that individuals always pursue that action which produces the greatest amount of happiness. But in so doing, the paradox suggests, this might prevent achieving other great sources of happiness. By seeking to maximize each token instance of happiness, the overall amount of happiness achieved in one’s life might in fact be lower than if one instead pursued other activities that were not necessarily the actions that, of all other alternatives at that time, were maximally conducive to happiness. Peter Railton suggests a way to circumvent the paradox by drawing the following distinctions:

---

Subjective hedonism is the view that one should adopt the hedonistic point of view in action, that is, that one should whenever possible attempt to determine which act seems most likely to contribute optimally to one’s happiness, and behave accordingly. Objective hedonism is the view that one should follow that course of action which would in fact most contribute to one’s happiness, even when this would involve not adopting the hedonistic point of view in action. An act will be called subjectively hedonistic if it is done from a hedonistic point of view; an act is objectively hedonistic if it is that act, of those available to the agent, which would most contribute to his happiness.\textsuperscript{61}

Railton continues:

Let us call someone a sophisticated hedonist if he aims to lead an objectively hedonistic life (that is, the happiest life available to him in the circumstances) and yet is not committed to subjective hedonism. Thus, within the limits of what is psychologically possible, a sophisticated hedonist is prepared to eschew the hedonistic point of view whenever taking this point of view conflicts with following an objectively hedonistic course of action. The so-called paradox of hedonism shows that there will be such conflicts: certain acts or courses of action may be objectively hedonistic only if not subjectively hedonistic. When things are put this way, it seems that the sophisticated

\textsuperscript{60} Peter Railton, “Alienation, Consequentialism, and the Demands of Morality,” \textit{Philosophy and Public Affairs} 13, no. 2 (Spring 1984): 134-171. Railton’s concern is ultimately to address the problem of alienation and morality—that is, (in his words): “whether being moral is a matter of taking a moral point of view and whether there is thus some sort of necessary connection between being moral and being alienated in a way detrimental to human flourishing (140).” Railton contends that one can remove the alienation problem from morality if we take morality to be a disposition that generally requires that one do x, but that when doing x conflicts with certain significant things, where this conflict might lead to alienation, doing x can be overridden.

\textsuperscript{61} Railton, “Alienation, Consequentialism, and the Demands of Morality,” 143.
hedonist faces a problem rather than a paradox: how to act in order to achieve maximum possible happiness if this is at times—or even often—*not* a matter of carrying out hedonistic deliberations.\(^62\)

This is strikingly similar to Gauthier’s claim regarding constrained maximization. A constrained maximizer will forgo taking up a maximizing conception of reason (and behaving as a straightforward maximizer) whenever doing so conflicts with making one’s life go as well as possible. Just as the paradox of hedonism shows that there will be conflicts between subjectively and objectively hedonistic acts, so too does the Prisoner’s Dilemma show that there is a conflict between maximizing and optimizing acts. And just as the sophisticated hedonist must then decide which of her options will lead to a maximization of happiness, so too does the rational person face a problem of how to act in order to obtain the utility-maximizing outcome.

Railton asks: “Is it bizarre, or contradictory, that being a sophisticated consequentialist may involve rejecting subjective consequentialism? After all, doesn’t any adherent of subjective consequentialism also seek to lead an objectively consequentialist life?”\(^63\) And answers: “He may, but then he is mistaken in thinking that this means he should always undertake a distinctively consequentialist deliberation when faced with a choice.”\(^64\) To illustrate this, he presents the following:

---


It is well known that in certain emergencies, the best outcome requires action so swift as to preclude consequentialist deliberation. Thus a sophisticated consequentialist has reason to inculcate in himself certain dispositions to act rapidly in obvious emergencies. The disposition is not a mere reflex, but a developed pattern of action deliberately acquired. A simple example, but it should dispel the air of paradox.\(^{65}\)

According to Railton, “people can learn to avoid certain characteristically self-defeating lines of thought…and the sophisticated consequentialist may learn that consequentialist deliberation is in a variety of cases self-defeating, so that other habits of thought should be cultivated.”\(^{66}\) The same, I contend, goes for the constrained maximizer. I thus conclude that we ought to grant to Gauthier that actions are rational when recommended by rational dispositions, even if those actions are not individually utility-maximizing (and are, in fact, out of equilibrium in one-shot Prisoner’s Dilemmas). We thus we arrive at a reconciliation between moral dispositions, the actions they recommend, and rationality.


CHAPTER SIX

Reconciling Reason, Evolution, and Morality

In this thesis I have taken up two main questions. The first is the descriptive question of how to explain the emergence of cooperation. The second is the normative question of why one should be moral (or cooperate). These two are structurally similar. The constraints imposed by natural selection in the descriptive case and by rationality in the normative case make explaining or justifying cooperation difficult. How is it possible that cooperation evolved, given the workings of natural selection? How can we justify cooperation, given the self-serving conception of rationality? Regarding the former, I have argued that cultural group selection provides us with the most plausible explanation of the emergence of the widespread cooperative behaviours among human beings. In the normative context, I have argued that David Gauthier’s argument for the rationality of adopting the disposition of constrained maximization is a defensible route to reconciling morality with rationality.

In what follows I will examine the relationship between the descriptive and normative projects. I will argue that the descriptive and normative projects are not only dependent on one another, but converge on the same outcome. This convergence comes from two independent lines of enquiry. Rationality requires that we form
dispositions towards constrained maximization. The coevolution of genes and culture has made us creatures with prosocial dispositions, which likewise constrain the pursuit of our interests. Thus we have a coincidence between what reason dictates and evolution produces.

**Summarizing Our Findings**

It will first be useful to take stock of what has been established thus far. Our descriptive story attempted to uncover the mechanisms responsible for the emergence of cooperation. We began, in chapter one, by looking at Skyrms’s employment of evolutionary game theory to explain the emergence of cooperation. I argued that EGT explanations come up short and cannot account for the tremendous stability observable in human social groups. Skyrms’s reliance on mechanisms that apply to mindless organisms fails to extend to models involving agents with higher levels of rationality. As I argued, an increased level of rationality increases the possibility of free riding and, thus, destabilizes cooperation.

In chapter two we reintroduced high levels of reason and turned to Binmore’s appeal to the folk theorem to explain cooperation. I argued that here too the door is left open to free rider strategies. Binmore’s model requires a punishment structure that either becomes implausible as group size increases, or raises a second-order free rider problem which calls for an explanation of why self-regarding individuals would be willing to punish defectors.
Given that cooperation is prevalent and free riders are not, there is reason to suppose that the above-mentioned mechanisms do not provide a complete explanation of cooperation. In chapter three I turned to Sober and Wilson’s group selection hypothesis, according to which cooperation emerged as a biological group-level adaptation. I argued that the notion that cooperation is such a group-level adaptation is intriguing, but that the conditions required for large-scale cooperation to emerge as an outcome of biological group selection are unlikely. Specifically, I argued that it is not plausible that human groups are naturally organized so as to yield the requisite variation between groups and homogeneity within groups to permit biological group selection. But the same, I argued, is not true at the cultural level and, thus, group selection on cultural variants promises to solve the cooperation problem in nature.

Chapter four took up the possibility of cultural group selection, as championed in the works of Richerson and Boyd. I argued that their hypothesis regarding the emergence of cooperation as a result of group selection on cultural variants is able to explain the stability of cooperation in a way consistent with facts about nature, while avoiding the pitfalls that earlier accounts encountered. I argued that if we take cultural group selection seriously, we arrive at an explanation of widespread cooperation that is compatible with evolutionary processes. Furthermore, the resulting environment helps to shape our evolved psychology to generate prosocial dispositions.

The first four chapters of the thesis thus culminate in the following conclusion: we cannot tell a story of the emergence of cooperation without invoking culture. Part of this story involves appealing to the presence of cooperative norms that
help make groups do better than non-cooperative groups, and create an environment where prosocial dispositions are selected via individual selection.

In the last chapter I turned to the justificatory project in ethics. Moral theories have traditionally asked how we ought to live. Social contract theories have generated answers to this question by asking which principles rational persons would agree to under idealized conditions. But, as we have seen, the employment of reason in the sense of rationality is problematic in that it leads to non-cooperative outcomes. Gauthier, we saw, argues for the rationality of constrained maximization, a disposition that, if employed by all would generate cooperative and mutually beneficial outcomes. I examined and defended Gauthier’s justification for the rationality of constrained maximization. I argued that the rationality of disposing oneself to constrained maximization depends on how well that strategy does against other strategies within a population and this, I argued further, depends primarily on the plausibility of translucency.

If we grant that translucency is plausible, as I have contended that we should, then adopting constrained maximization will permit individuals to reach a cooperative outcome. This cooperative outcome yields a higher utility than does mutual defection but imposes the constraint that individuals refrain from pursuing what individual utility-maximization would dictate—viz., unilateral defection. We saw that establishing the rationality of constrained maximization thus requires a reconceptualization of rational action to strive towards mutually-advantageous outcomes rather than independently maximizing acts. I argued that there is no clear
rationale why this cannot be done and, thus, no clear argument against the rationality of constrained maximization. And insofar as constrained maximization is a moral disposition, the result is a reconciliation between morality and rationality.¹

Situating the Descriptive and Normative

All this leads directly to the question of how the descriptive and normative projects are connected. On this there is some controversy, primarily generated by critics’ resistance to a purely evolutionary ethics. One of the more common lines of criticism facing approaches to empirically informed—and more specifically in this case, evolutionarily informed—ethics is the “is/ought” problem. Hume famously raised this problem as follows:

In every system of morality which I have hitherto met with, I have always remark’d, the author proceeds for some time in the ordinary way of reasoning, and…makes observations concerning human affairs; when all of a sudden I am surpriz’d to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought or an ought not. This change is imperceptible; but is, however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, ‘tis necessary that it shou’d be observ’d and explain’d; and at the same time that a reason should be given, for what seems altogether

¹ Recall that I use the terms “morality” and “cooperation” here interchangeably. There are moral norms that do not have to do with cooperation, but I am not concerned with these.
inconceivable, how this new relation can be a deduction from others, which are entirely different from it.²

In short, Hume’s objection (commonly referred to as “Hume’s Law”) runs as follows: “is” is one thing; “ought” is another; and we cannot derive the latter from the former. That is, we cannot derive any normative guidance from facts about nature alone. Accordingly, if we take Hume’s Law seriously, then we will not be able to conclude, from the fact that evolution has generated cooperation and dispositions toward social behaviour that we ought to cultivate those dispositions. Thus, to achieve any normative guidance about what we ought to do will require supplementing the descriptive account with something else. We will require a substantive normative moral theory.

There is, however, room to object to the view that a proper or full picture of the nature of morality and its prescriptions will require the employment of a normative moral theory. We saw in the introduction that Binmore thinks that by reducing morality to evolutionary terms we eliminate any need to ask about what we ought to do. On Binmore’s view, since our attitudes and moral rules are produced by evolutionary processes, we have no control over them. Thus, it makes no sense to raise the question of how we ought to act or what dispositions we ought to form.

The descriptive approach can be seen as a response to problems with the normative branch of social contract theory. One central motivation for undertaking the descriptive enterprise, we have seen, stems from recognition of some of the

limitations of the *homo economicus* model of human action. This model stands at the heart of most economic, political and moral theories, and provides the foundation for the contractarian arm of the social contract tradition. With this model comes a particular conception of reason, which, as we have seen, runs us into problems with issues of compliance. As long as we hew to that conception of reason and deploy it as Hobbes does, we will not be able to either justify or explain cooperation.

Binmore claims that we can bypass this difficulty by simply replacing reason with evolutionary processes. Rather than asking how we ought to live, we simply need to identify the processes that permit us to solve cooperation problems. We can then make use of this knowledge in order to make reforms to the social contract.³ Binmore says:

> The moral rules that govern our behavior consist of a mixture of instincts, customs, and conventions that are simultaneously more mundane and more complex than traditional scholarship is willing to credit. They are shaped largely by evolutionary forces—social as well as biological. If one wishes to study such rules, it doesn’t help to ask how they advance the Good or preserve the Right. One must ask instead how they evolved and why they survive. That is to say, we need to treat morality as a science.⁴

We have seen that Binmore turns to evolution to solve the equilibrium selection problem. A solution to this problem constitutes a selection of the principles (or social

---

³ As Skyrms says, “those who wish to change the world had better first learn to describe it.” *Evolution of the Social Contract*, 109.

norms) according to which a society will govern itself. These include moral norms. Social norms concern the standards a society has for how people ought to act. These “oughts” can range over aesthetics, prudence, and morality. That is, individuals can act in ugly ways, ways that hurt themselves, and morally wrong ways. The sphere of morality will include things such as lying, killing, stealing, and so on. It also concerns keeping one’s bargains, i.e., cooperating, which is my concern in this thesis. If social norms are mere products of evolution then, so the argument goes, there is no need for reasoning about which norms ought to be adopted. However, as I shall now argue, if the story that I have told is plausible, the descriptive approach does not eliminate these normative questions, but rather underlines their importance.

The Need for the Normative

We have seen that two of the central questions in moral theory are: (1) what is the source of moral norms? and (2) why should anyone comply with them? I have argued that individual selection based models and biological group selection models do not provide adequate accounts of the emergence of large-scale human cooperation. They fail either because they cannot explain the scarcity of free riding, or because the background conditions required by them are unlikely to have obtained in our evolutionary history. I argued, in chapter four, that cultural group selection, as described by Richerson and Boyd, explains what other theories could not.

I will use the terms “moral norms” and “cooperative norms” interchangeably. I will do the same with “social norms” and “cultural norms.” Moral/cooperative norms are a subset of social/cultural norms, which include other norms such as aesthetic and prudential norms. The moral norms that I am concerned with in what follows are those of cooperation.
If we accept this story, then the emerging picture of moral norms is this: they are cultural variants that have been selected on the basis of the benefits that they confer onto groups that possess them. We saw that different groups operate with different social norms. This variation between groups, we have seen (in addition to other factors), sets the stage for group selection. Groups adhering to more successful norms outcompete other groups, and the more successful norms will flourish. Among the norms that contribute to a group’s success are cooperative norms. Thus, certain norms exist because they permitted groups operating with them to outcompete groups operating with different norms. We saw, further, that the reasons for compliance with norms emerge partly through genetically endowed prosocial dispositions producing a propensity towards norm-conformity, and partly through a socialization process and sensitivity to cultural rules backed by internal and external sanctions.

Thus, cooperative norms make groups do better. But they are cultural artifacts and are therefore subject to the same transmission mechanisms as other cultural variants. These transmission mechanisms—as well as the content of cultural norms—are constrained by our evolved psychology and are dependent on certain biases as well as reasoned reflection. Cultural transmission mechanisms are different from those involved in the propagation of genes. In the case of genes we have no say over what we get and pass on. Natural selection works on random mutation as a “culling process” that eliminates unsuccessful variants. In the case of culture, however, we do. We have seen that there are certain transmission mechanisms in place that help to guide what kinds of cultural variants get transmitted. While cultural transmission
sometimes occurs by random variation and natural selection, cultural variants are also transmitted by biased transmission and guided variation. These transmission mechanisms depend on an agent’s preferences and desires. We may not agree with certain social norms and decline to adopt them. Or we may modify the norms and pass them along in that modified form. In short, reason and reflection play a role in cultural transmission in a way that they do not play one in genetic transmission. Since genetic variations occur blindly by chance, and cultural variations are guided, the evolutionary explanations are different. As cultural variants, the content of social norms will thus depend on human reason. On this, Sober and Wilson say: “Rational thought may account for many phenotypic differences among groups. Among anthropologists, it has not been accorded the attention it deserves as an agent of cultural change.” The benefits of cooperative norms help them get situated within a population. But whether they are adopted in the first place depends on others’ reasons for doing so. The content of the norms, and whether they survive will depend in part on whether others are willing to adopt them. And that will depend on what they take to be good reasons to adopt them. Thus the return of the normative insofar as human reason plays a role in the content and transmission of moral norms.

But this is not the return of pure unadulterated reason. As Richerson and Boyd say, “our innate social psychology furnishes the building blocks for the evolution of complex social systems, while simultaneously constraining the shape of these

---


7 Sober and Wilson, *Unto Others*, 153.
Norms may be culturally transmitted, in that they are not genetic and are modifiable. But they might not be easily changed. What gets transmitted is constrained by our evolved psychology. Given these constraints, there is only so much that can be done to change the norms that we arrive at.

I thus conclude that, since human reason plays an uneliminable role in the origin and transmission of cooperative norms, descriptive (in this case evolutionary) approaches cannot replace normative theories. I now turn to my last aim of the chapter, namely showing the convergence between the outcomes of the specific descriptive and normative projects that I have defended.

**Point of Convergence**

I have argued that group selection on cultural variants provides a plausible explanation of the evolution of cooperation among human beings. Cultural group selection permits us to explain the emergence of behaviours that are at odds with selective processes at the individual level but that are beneficial to groups of individuals that display these behaviours. Explaining the possibility of cooperation in this context requires us to shift our perspective from the level of the individual to the level of groups. Individually, cooperation is more costly than selfishness. Collectively, cooperation pays.

A similar shift in perspective is required to justify the rationality of cooperation. Gauthier’s project is to show that moral behaviour is rational, in spite of

---

it being disadvantageous at particular instances. This parallels the explanation of cooperation in nature, which, as I have argued, evolves in spite of it being apparently disadvantageous at the individual fitness level. Non-cooperation is utility-maximizing at the level of outcomes: if rationality is evaluated as a best reply to one’s partner’s actions, non-cooperation will always be rational. But by shifting the evaluation of rationality from outcomes to strategies, the cooperative—and superior—outcome can be achieved. This perspective permits us to rationally justify the constraints that morality requires.

Thus, just as it is the case that, contrary to what we might expect, evolution supports cooperation, so too is it the case that, contrary to appearances, rationality supports cooperation. There is an advantage to cooperative behaviour in a particular context. That is, when others cooperate and when cooperation permits outcompetition of other groups in the evolutionary context, and when it permits a mutually preferred outcome to universal defection in the moral context, then cooperation is better than non-cooperation.

If the correct explanation of the existence of cooperation in nature appeals to group selection on cultural variants, then we also are able to arrive at an evolutionary story of the emergence of certain prosocial dispositions—viz., ones that dispose us to comply with norms and agreements. I argued in chapter four that group selection helps to shape an environment that favors prosocial psychological mechanisms, which in turn help to stabilize cooperation. In chapter four I also suggested that we might tell the story of the transition from culturally driven behaviour to genetically-
dictated behaviour in terms of Baldwinian evolution: that is, that the culturally
generated environment created new selection pressures that favored the genetic
evolution of prosocial dispositions. Thus, what starts off as a cultural phenomenon
transforms to a genetic disposition. Assessing the viability of this story lies beyond
the scope of my aim in this thesis, and there is no settled opinion on it in the
literature. What can be said concretely, however, is that prosocial dispositions play a
role in the emergence and maintenance of cooperation.

Thus, group selection helps to shape an environment that favors prosocial
psychological dispositions. Prosocial dispositions are ones that, if evaluated at
individual instances, are fitness-decreasing (or at least are not fitness-maximizing).
And insofar as this is true, they are also structurally similar to those dispositions
required by constrained maximization. And if Gauthier is right that the dispositions
recommended by constrained maximization are rationally defensible, then we will
have shown that the dispositions emerging from the evolutionary story that I endorse
can also be rationally defended.

There is an advantage to cooperative behaviour in a particular context. That is,
when others cooperate and when cooperation permits outcompetition of other groups
in the evolutionary context, and when it permits a mutually preferred outcome to
universal defection, then cooperation is better than non-cooperation. Both constrained
maximization and evolved prosocial dispositions (e.g., empathy, norm-conformity,
etc.) are dispositions that, if taken at face value, might be fitness-decreasing.
Nonetheless, they are when properly understood in the contexts in which we find
them, advantageous. Prosocial emotions are advantageous in the new cultural environment. Constrained maximization is advantageous in an environment where translucency is common.

If others are constrained maximizers, then it becomes rational to likewise so dispose oneself, provided one’s dispositions are reliably detectable. Richerson and Boyd say:

People are likely to discriminate against genotypes that are incapable of conforming to cultural norms. People who cannot control their self-serving aggression ended up exiled or executed in small-scale societies and imprisoned in contemporary ones. People whose social skills embarrass their families will have a hard time attracting mates. Of course, selfish and nepotistic impulses were never entirely suppressed; our genetically transmitted evolved psychology shapes human cultures, and as a result cultural adaptations often still serve the ancient imperatives of inclusive genetic fitness. However, cultural evolution also creates new selective environments that build cultural imperatives into our genes.\(^9\)

McClenne\(\)n says further that “those who are capable of rule-governed choice do better in terms of furthering whatever interests they have than those who are not, and this has direct bearing on the credibility of claims about what counts as a rational approach to various forms of interaction.”\(^{10}\)

---


\(^{10}\) McClenne\(\), “The Strategy of Cooperation,” 204.
We can end by noticing that both Gauthier and Rawls appeal to evolution to bolster their accounts of justice. Gauthier notices the potential an evolutionary analysis can have to strengthen his defense of constrained maximization. He writes:

[Let us note an interesting parallel to our theory of constrained maximization—Robert Trivers’ evolutionary theory of reciprocal altruism. We have claimed that a population of constrained maximizers would be rationally stable; no one would have reason to dispose herself to straightforward maximization. Similarly, if we think of constrained and straightforward maximization as parallel to genetic tendencies to reciprocal altruism and egoism, a population of reciprocal altruists would be genetically stable; a mutant egoist would be at an evolutionary disadvantage. Since she would not reciprocate, she would find herself excluded from co-operative relationships.]

Rawls too appealed to evolution to provide a basis for the sense of justice required by his contractualism:

In arguing for the greater stability of the principles of justice I have assumed that certain psychological laws are true, or approximately so….We may note however that one might ask how it is that human beings have acquired a nature described by these psychological principles. The theory of evolution would suggest that it is the outcome of natural selection; the capacity for a sense of justice and the moral feelings is an adaptation of mankind to its place in nature….It seems clear that for members of a species which lives in stable social groups, the ability to comply with fair cooperative arrangements and to

---

develop the sentiments necessary to support them is highly advantageous, especially when individuals have a long life and are dependent on one another. These conditions guarantee innumerable occasions when mutual justice consistently adhered to is beneficial to all parties.\footnote{John Rawls, \textit{A Theory of Justice}, 502-503.}

In the same spirit, if the story that I have told is right, cultural group selection provides us with an evolutionary process that will favour dispositions resembling constrained maximization, and constrained maximization enables us to harmonize morality and rationality. And if so, evolution gets us closer still to a reconciliation between morality and rationality.
Bibliography


Bell, Adrian V. “Why Cultural and Genetic Group Selection are Unequal Partners in The Evolution of Human Behavior.” *Communicative and Integrative Biology* 3, no. 2 (March/April 2010): 159-161.


_____.


_____.


