UNIVERSITY OF CALIFORNIA,
IRVINE

Diversity and Cooperation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

Justin Pearce Bruner

Dissertation Committee:
Distinguished Professor Brian Skyrms, Chair
Associate Professor Simon Huttegger
Associate Professor Peter Vanderschraaf

2014

DEDICATION

To Gary, Julie, Jill, Sarah and Roy

All pluralists, in their own way.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Acknowledgments While at the University of California, Irvine, I have been supported by a Social Science Merit Fellowship from the School of Social Science. I was supported by a National Science Foundation Grant (No. EF 1038456) administered by Simon Huttegger in the Summer term of 2012, as well as the Winter, Spring, Summer and Fall terms of 2013. I was supported by a generous Dean's fellowship in the Winter of 2014. I am very grateful to all of these funding sources.

I've benefited tremendously from presenting the contents of this dissertation at a number of conferences. I received valuable feedback on all substantive chapters of this dissertation from the Social Dynamics Seminar co-taught by Brian Skyrms, Louis Narens and Don Saari. I also benefited from feedback from the Institute for Mathematical Behavioral Sciences Graduate Student Conference 2012 at UC Irvine, Formal Ethics 2012 hosted by the Center for Mathematical Philosophy in Munich, Philosophy of Biology in the Desert at Arizona State University, and the Association for the Study of Religion, Economics and Culture at Chapman University.

My excellent dissertation committee members deserve many thanks. I owe a great deal to Brian Skyrms. I benefited not only from his direct guidance, but also from the department and school he helped craft. During my time at UC Irvine, there were many exciting conferences, interesting talks and engaging new people to meet. This in large part was his doing. Simon Huttegger was immensely helpful throughout all stages of graduate school and was generous with both his time and funds. I am lucky to have found such a great mentor and co-author early in my career. Peter Vanderschraaf was kind enough to sit on my committee even though most of our interactions up to this point have been via email. His work combining political philosophy and game theory has had a large impact on me, as evidenced by the contents of this dissertation.

In addition to my committee members, many of the other professors and graduate students have helped me thrive here at UC Irvine. Elliott Wagner generously helped guide me through my first year of graduate school. I would also like to thank the many other friends and graduate students who I benefited from intellectually: Cailin O'Connor, Hannah Rubin, Bennett Holman, Nathan Fulton, Greg McWhirter, Abraham Morris, Skyler Nelson, Jason Messer, Jim Weatherall, Ethan Galebach, Ben Feintzeig, Sarita Rosenstock, Michael Caldara, Ryan Kendall and Schatz. I've learned much from professors Louis Narens, Don Saari, Michael McBride, Kyle Stanford, Jeffrey Barrett, and Jean-Paul Carvalho.

Some friends and teachers from before my time in graduate school deserve mention as well. In particular, Lara Buchak helped me realize the potential game theory has in philosophy and encouraged me to pursue my interests in graduate school. In many ways, this dissertation is an extension of my undergraduate thesis that she supervised. Cara McGraw helped nurture my philosophical interests early on, and continues to be a good friend and engaging interlocutor.

Finally, my parents deserve many thanks. They have always been supportive, even when I decided to pursue philosophy professionally.

CURRICULUM VITAE

Justin Pearce Bruner

2010 B.A. Economics, University of California, Berkeley

2014 Ph.D. Philosophy, University of California, Irvine

ABSTRACT OF THE THESIS

Diversity and Cooperation

By

Justin Pearce Bruner

Doctor of Philosophy in Philosophy

University of California, Irvine, 2014

Distinguished Professor Brian Skyrms, Chair

The present dissertation is an exploration of the effect of diversity on social contract formation and the evolution of cooperation. This work stems from the pioneering efforts of economist Arthur Robson, who first explored the role of costless pre-game communication in strategic interactions. When communication is permitted, individuals playing a game can condition their behavior on the signal received from their counterpart. For my purposes, I interpret these signals as racial markers or cultural identifiers, which in turn provides a formal framework to precisely study a number of issues relevant to political and social philosophy.

My first substantive chapter, "Diversity, Tolerance and the Social Contract," starts by formalizing the state of nature as a game in which individuals can either choose to remain in the state of nature or attempt to found a social contract. I assume there exists some natural diversity in the population, and that individuals are pre-disposed to behave cooperatively with those who are more similar to themselves. I uncover an interesting relationship between diversity, tolerance and the social contract. Social contract formation is possible but initially comes with a cost for both diversity and tolerance. That is to say, individuals quickly all adopt the same signal and only behave cooperatively with those who send similar signals. This, however, is not a long-

term feature of the population. In the long run, individuals slowly become more tolerant, cooperating with those who are quite dissimilar to themselves. The circle of cooperation expands, and soon all can partake in a thriving social contract.

My second substantive chapter, "Racists and Minorities in Population Games," focuses on the welfare of racial minorities, as well as explores one means of expunging racist attitudes and behaviors from a population. I show that in a wide range of games, minorities are at a distinct disadvantage. Consider the Nash demand game, a canonical bargaining game in which a resource is to be divided between two individuals. I show that in this game, minority status translates into a bargaining disadvantage. In other words, the population tends to settle on an equilibrium in which individuals from the racial majority receive the bulk of the resource. Interestingly, this minority disadvantage is not due to differential abilities or effort, but is instead simply in virtue of the minority's relative size. Second, I consider one means of reducing racist behavior. If individuals are allowed to send a plastic signal that is independent of their fixed racial signal, then individuals tend to condition their behavior on the plastic signal of their counterpart, which in turn facilitates high levels of cooperation.

My final substantive chapter, "The Possibility of Pluralism," explores cooperation and diversity in the context of a liberal pluralistic society. In such a society, many different valid conceptions of the good would exist, and individuals would ideally be tolerant of different moral beliefs and practices. Yet under what conditions is such an arrangement possible? Taking my cue from the political philosopher Gregory Kavka, I investigate how disagreement among individuals with different value systems would be settled. Individuals can either compromise and find some middle ground, or dig their heels in and refuse to concede. Using computer

simulations, I identify that conflict is minimized when, among other things, individuals are embedded on a social network and are allowed to employ somewhat sophisticated strategies, such as tit-for-tat.

Chapter 1

INTRODUCTION

How is cooperation possible in a Darwinian world of survival of the fittest? This question has been entertained by philosophers, biologists and social scientists for decades now. The answer that has emerged is, not too surprisingly, that cooperation is possible in certain circumstances, and unattainable in others. It is not always the case that agents will be able to interact with the same individual repeatedly, nor is there a guarantee that reputations can be easily tracked. Conflict between groups may not be intense enough to support high levels of so-called strong reciprocitors. Devices that help correlate behavior may not be on offer. And so on.

Yet while most work on the evolution of cooperation attempts to identify social mechanisms that promote pro-social behavior, surprisingly little effort has been spent on better understanding the ways in which cooperation can fail to arise.[1] What impediments are there to pro-social behavior, and how can they be overcome? One natural suggestion is to investigate the role diversity plays in strategic settings. While the problem of cooperation may be a difficult one, these difficulties are compounded in the presence of rampant ethnic, racial or religious diversity.

Not surprisingly, political philosophers have spilt much ink discussing such settings. Religious or ethnic diversity can lead to outright conflict, and such clashes can often have large, catastrophic effects on society. For many social thinkers, such episodes of violence and in-fighting was to be avoided at all costs. Thomas Hobbes, for instance, was quite influenced by the chaos of the English Civil War. Avoiding conflict incited by racial, ethnic, or in this case, religious, diversity was a high priority for Hobbes, and one supposed virtue of his political philosophy outlined in *Leviathan* is that an absolute sovereign can easily squelch such civil unrest. Placing such a premium on stability entails that certain means of establishing peace in a diverse society, such as a modus vivendi, are ultimately unsatisfactory.

Concerns about diversity, cooperation and social cohesion continue to this day, attracting the attention of philosophical heavyweights. Peter Singer, contra Hobbes, suggests that

[1] See Smead and Forber (forthcoming), who demonstrate that cooperate has difficulty getting off the ground in the prisoner's delight in a finite population.

cooperation between various religious, ethnic and racial groups – and even between different species – will naturally emerge. The "circle of cooperation" will slowly, but inevitably, expand, and in the long run, all can participate in a thriving social contract. Relatedly, Rawls has spent a career better characterizing and defending his vision of a pluralistic, liberal society. In such a community agents subscribe to different so-called comprehensive doctrines, and stability is possible because each agent can justify the state from within the confines of her comprehensive doctrine.

In the course of this dissertation I investigate the effect of diversity in strategic settings, in addition to exploring a number of different social mechanisms that can engender high levels of cooperation in a diverse population. We will see that social contracts, fair division and even political pluralism can be attained in a variety of different circumstances. In this sense this dissertation is a natural extension of the laudable project of natural social contract theory outlined first by Hume, and continued to this day by both philosophically minded economists and naturalistic philosophers. The dissertation will for the most part rely on computer simulations and the tools of evolutionary game theory, discussed more below. This will enable us to paint a vivid picture of how social contracts unfold in the face of persistent diversity. It will also provide us a means of precisely assessing some famous philosophical claims, regarding, among other things, the possibility of peaceful cooperation in the state of nature, as well as the stability of a tolerant and pluralistic society.

We interpret diversity broadly enough, and assume it can refer to racial or ethnic diversity, as well as other, more mundane, senses of the word. Diversity can also refer to one's belief system, conception of the good, or value system. How cooperation is possible is a pressing question, and findings from both social psychology and experimental economics highlight how difficult pro-social behavior is to sustain in the face of diversity. Glaeser et al. (2000), for example, find that in trust games agents behave less cooperatively when paired with a member of a different race. Similarly, Krupp et al. (2008) show that in experimental settings people contribute less to a public good when the other members are less physically similar to themselves. Evidence exists outside the lab as well, as Miguel et al. (2005) document that public good provisions in Africa were significantly lower in regions with higher ethnolinguistic diversity.

One way of nicely approaching this subject matter is to utilize the formal framework provided by the mathematical theory of games, i.e., game theory. Although game theory is relatively new (developed in the 1940's and 1950's) it has been used as a tool by philosophers for quite some time, going back to Richard Braithwaite's (1955) initial call to incorporate game theoretic analysis in moral theory. Many game theorists have also noted the natural overlap between the two disciplines. Roger Myerson, for example, notes that "modern game theory is the modern continuation of Hobbes's great work, as theoretical physics is of Newton's."[2] Even more promising, is evolutionary game theory, which has already helped shed light on a number of distinct sub-fields of philosophy, such as logical inference, scientific theory choice, social epistemology, language, and of course, social and political philosophy.

In the realm of political philosophy, game theory has long been utilized to help formalize and represent strategic interactions in the state of nature. David Gauthier and Gregory Kavka, as well as Ed Curley, have all argued that Hobbes' state of nature can be formalized as a simple two-person simultaneous move game. Robert Sugden (1986), Kenneth Binmore (1994) and Brian Skyrms (1996) have together made the "dynamic turn," moving away from the static framework initially proposed by classical game theorists to the evolutionary game theory developed in the 1970s by John Maynard Smith, among others. This move is essential, for it not only allows us to identify stable arrangements, but provides us with a means of assessing which of the available outcomes on offer are more likely to emerge (for a similar methodological discussion of the use of game theory in philosophy and the sciences, see Huttegger and Zollman, 2012). This dissertation can be seen as an extension of this cultural-evolutionary work.

In general, we consider strategic situations that are pertinent to political philosophy, and attempt to determine how behavior will unfold in such circumstances. The exact analysis differs depending on the game. For example, in the prisoner's dilemma, where there is only one pure strategy Nash equilibrium, focus is on what sort of additional social mechanisms can lead to out of equilibrium behavior – i.e., how can cooperation be sustained? In other games, such as the stag hunt of chapter 2, there are two possible pure-strategy equilibria. In this case our attention turns to which of these possible equilibria is the more likely outcome of an evolutionary process. This

[2] Quotation comes from the back cover of Binmore, 1994.

involves introducing some tools, such as imitative dynamics and the replicator equations. Additionally, computer simulations are employed to investigate the behavior of a number of agent-based models. All of these tools are used together to help provide a fuller picture of the phenomena in question.

One vital theme that underscores all chapters of the dissertation is the important role of communication in games. In particular, we look at one form of pre-play communication, so-called cheap talk. Cheap talk refers to communication between agents that is not costly. Signals can be freely sent, meaning the act of signaling neither lowers the agent's utility or fitness. Of course, since signals have no cost attached to them, signals cannot function as screening mechanisms (as they do in models considered by Michael Spence and Alan Grafen). Since I have incentive to lie, no one would believe me if I simply *claimed* to be an efficient worker. Cheap talk cannot convince in such a setting. What, then, can cheap talk accomplish? It is easy to see how cheap talk can be of use in pure coordination games in which there are multiple equilibria. With neither agent incentivized to lie, pre-game communication can be effectively utilized to coordinate behavior. Yet does cheap talk have any bite outside this narrow context? One relatively recent realization has been the immense importance of cheap talk in non-trivial games. Arthur Robson (Robson, 1990) was one of the first to realize the importance of cheap talk in evolutionary contexts, arguing that cheap talk destabilizes the all defect equilibrium in the prisoner's dilemma. We will see that cheap talk has a large role to play, and can help coordinate behavior, among other things. The exact effect cheap talk has on a game can only be determined by examining the details, but one thing is certain – costless pre-game communication does matter across a variety of different games, and this is an important point that has gone somewhat unappreciated in the social sciences.

The dissertation is organized as follows. We first begin with an exploration of the stag hunt. It has been suggested that the stag hunt nicely captures the strategic situation of individuals in the state of nature. Thus the stag hunt provides the perfect setting to explore the effect diversity has on social contract formation. The results are striking. We see that in the short run diversity is expelled from the population, as all adopt similar traits. This homogeneity in turn ensures high levels of cooperation. Yet this is not a long term feature of the population. In the long run two related things occur. First, the average tolerance level of the population increases. Secondly, as the tolerance level increases, more trait mutants can thrive and participate in the stag hunt. Thus

chapter 2 uncovers an interesting moral progression, where, starting from the state of nature, we transition from an exclusive social contract to an increasingly more tolerant society. Intolerance and homogeneity play the important role of establishing the contract, but are not long term features of the community.

This partly vindicates Singer. The circle of cooperation expands, once cooperative behavior is established. Yet inciting collective action is really only half of a social contract. The second part, which has been overlooked by Singer, pertains to how the *fruits of cooperation are divided* (see Skyrms, 2013 and Wagner, 2012). In other words, once two agents cooperate to bring about some resource, they must then determine how to allocate the good. Inequalities can easily emerge in a decentralized community of agents. Chapter 3 explores such bargaining scenarios, and in particular, we investigate a model in which agents must interact in a simple bargaining game, the mini-Nash demand game. Axtell et al. (2000) have shown that in such a game, if individuals are given permanent tags and are allowed to condition their behavior on the tags of their counterpart, inequalities can naturally emerge, and these inequalities track tags. We explore a similar model and see that when one group is substantially smaller than the other (i.e., when there exists a minority), these hierarchical arrangements are all the more likely, and members of the minority frequently receive the small slice of the communal pie. Finally, we briefly turn our attention to one means of reducing these tag-based strategies.

Lastly, we consider a different sort of diversity. In chapter 4 we attempt to determine under what circumstances agents who all subscribe to different conceptions of the good can peacefully coexist. An early answer was given by Gregory Kavka, who used some elementary game theory and evolutionary theory to argue that intolerant agents will, without severe governmental interventions, invade and take over a population of tolerant agents. This result is due primarily to the choice of game. We argue that instead of a prisoner's dilemma, a conflictual coordination game is more appropriate, and discover that in this setting, higher levels of tolerance are attainable. In particular, when agents are embedded on a social network and can utilize tit-for-tat, more tolerant and compromising dispositions can thrive. Finally, we argue that both Locke and Nozick's description of the state of nature can nicely be captured by a conflictual coordination game, and argue that under many circumstances peaceful settlement can be attainable without any governmental intervention or allegiance to a so-called protection agency.

Chapter 2

DIVERSITY, TOLERANCE AND THE SOCIAL CONTRACT

2.1 Introduction

Does diversity hinder the formation of an efficient social contract? If individuals do not take into account their differences when deciding whether to cooperate, then the answer is "no". Unfortunately, this is rarely the case. It is not uncommon for humans and other organisms to condition their behavior on how similar they are to those with whom they interact. This tendency manifests itself in the choices made in strategic situations, such as social contract games.

Stag hunt games are idealizations of social contract formation. In recent years philosophers and social scientists have attempted to assess under what conditions social contract formation is possible by appealing to evolutionary game theory and agent-based modeling. According to much of the literature on the evolution of cooperation, different social mechanisms can make the formation of a social contract more or less feasible.[3] Specifically, Skyrms (2004) demonstrates that if individuals are embedded in a social network and employ an imitate-the-best update rule, a social contract can be secured. Zollman (2005) extends this work by allowing individuals to send costless pre-game signals to their neighbors before engaging in a stag hunt. Both Skyrms and Zollman demonstrate that the likelihood of a social contract dramatically increases when simple and realistic social mechanisms are taken into account.

So-called "similarity-based strategies" are another type of realistic social mechanism. Individuals possess a number of observable traits and condition their behavior in a game on how similar they are to their counterpart. In the social realm, similarity-based cooperation has been documented numerous times. Individuals in both natural and experimental settings seem to condition their behavior on how similar they are to their counterparts. Glaeser et al. (2000), for example, find little cooperation in trust games when the players are of different races. Strikingly, Krupp et al. (2008) find that individuals are more likely to contribute to a public good the more

[3] For prime examples, see Axelrod (1984), Pollack (1989). For an in-depth but partial overview of the evolution of cooperation research, see Nowak (2006).

they physically resemble their fellow group members. These results are not limited to the laboratory. A number of natural experiments suggest that economic agents in real-life environments tend to employ similarity-based strategies. Miguel et al. (2005) document that regions in Africa characterized by high ethno-linguistic diversity tend to invest less in infrastructure and public goods.

The above examples seem to support a rather pessimistic story – cooperation seems possible but often comes at the price of diversity. Trust and aid are not extended to those who are too different from oneself. This leads to a natural question: is it possible to transform an intolerant group into an open and tolerant one? Peter Singer (1981) suggests it is possible for such a transition to naturally occur. Over time, Singer contends, the circle of cooperation will slowly expand, thereby permitting increasingly distinct members to join in the cooperative enterprise. We observe a similar dynamic in this paper in section 6.

Interest in tolerance is not restricted to moral and political philosophers alone. Biologists and social scientists alike seek to better understand the means by which tolerant behavior can be brought about. Milton Friedman, for example, in *Capitalism and Freedom* argues that a competitive exchange economy will over time weed out discriminatory practices in the business community. Companies that take into account the attributes of others that are impertinent from an economic point of view will be slowly driven out of business by competitors who only base their decision on economic values. In Friedman's view, there are strong self-interested reasons that compel people to become more tolerant. Friedman's narrative is in line with the findings of this paper. As we'll soon see, once a social contract is successfully established individuals have self-interested reasons to slowly become increasingly more tolerant.

Biologists and biological anthropologists have also realized the importance of uncovering mechanisms that promote tolerance. Unfortunately, most of the work by social scientists and theoretical biologists has instead been preoccupied with the task of *explaining the emergence* of ethnocentric behavior, not exploring means of reducing it,[4] with a few exceptions.[5] Smith and Szathmary (2000) note that while ethnocentric behavior has the upshot of promoting group

[4] See, for example, Axelrod and Hammond (2006).

[5] Muldoon et al. (2012) and Grim et al. (2005).

cohesion and high levels of cooperation *within* communities, it nonetheless leads to ethnic conflicts and violence between groups. While they do not offer any concrete recommendations, they encourage others to study different means of reducing the levels of ethnocentrism.

The scope of this paper is limited to investigating the effect similarity-based strategies have on the formation of a social contract. We will rely on a model loosely based on the collaborative work of Riolo et al. (2001).[6] In their model agents are endowed with a trait represented as a number in the interval from zero to one. Individuals cooperate if the distance between themselves and their partner in trait-space is less than their "tolerance level".

We follow the lead of Brian Skyrms and take the stag hunt game to capture the strategic considerations individuals face when determining whether to form or revise a social contract (more about this in the next section). Like both Skyrms and Kenneth Binmore, we are primarily concerned with how individuals in a community can coordinate their actions with one another. A stable arrangement which allows individuals to successfully coordinate with others will be referred to as a social contract. We for simplicity, restrict our attention to cases where there are only dyadic interactions – i.e., social contract games are two person games. While none of the major figures in the social contract tradition conceived of social contracts as attaining between just two individuals, many modern game-theoretic approaches to the social contract make this idealizing assumption. Binmore (2005), for example, models the social contract as a two person bargaining game.[7] If we conceive of a social contract as a stable arrangement that facilitates the coordination of individual actions in a society (as Skyrms and Binmore do), then investigating two-person games is a reasonable starting place.

[6] There are a number of theoretical models that attempt to demonstrate similarity-based strategies lead to high levels of cooperation in the prisoner's dilemma. While many of these findings appear promising, positive results often hinge on a few rather extreme assumptions implicit in the underlying models. We'll see that cooperation in the stag hunt does not require such strict assumptions. See Gilbert Roberts and Thomas Sherratt (2002) for a detailed criticism of Riolo et al.

[7] For further examples see Skyrms (1996) and of course Skyrms (2004).

We'll find that in a number of situations, attaining the pareto-dominant equilibrium in the stag hunt game is possible even though agents employ similarity-based strategies. If agents can change their traits with ease, cooperation is almost guaranteed. If traits are difficult to imitate because they correspond to permanent or semi-permanent features of the individual, such as race or culture, then social contract formation is improbable. In the case where traits are easily adopted, the population will naturally cluster in trait-space. These individuals will be rather intolerant and refuse to cooperate with those much different from themselves. Nonetheless, this clustering and intolerance allows for the formation of a social contract—all will be hunting stag. Furthermore, we'll see that this clustering in trait-space is not a permanent feature of the population. Due to experimentation, individuals slowly become more tolerant and over time are willing to cooperate with increasingly larger areas of trait-space. Diversity and the social contract seem completely compatible.

2.2 The Stag Hunt

The stag hunt, shown in Table 2.1, is a game between two players. Each has the option of hunting stag or hare. If both individuals opt to hunt stag the operation is a success and they receive a large bounty. If one of them opts to hunt hare, the stag hunt fails and the lone stag hunter is left empty handed. Hunting hare results in a small reward, but said reward is not contingent on the counterpart's behavior. Hunting stag may result in a large reward but is inherently risky. When S > H > V the two pure equilibria of the game are <stag, stag> and <hare, hare>.[8]

	Stag	Hare
Stag	S, S	V, H
Hare	H, V	H, H

Table 2.1: Normal form of the stag hunt with S> H > V.

This game is a stark contrast to the more popular prisoner's dilemma, shown in Table 2.2. While the stag hunt requires one to weigh the associated risks and benefits of hunting stag, it is

[8] In game-theoretic terms the <stag, stag> equilibrium is pareto dominant and the <hare, hare> equilibrium is considered risk dominant when H > .5S > 0. There also exists an unstable mixed Nash equilibrium.

always prudent to defect in the prisoner's dilemma. Defection is a strictly dominant strategy and hence is rational to perform regardless of the counterpart's behavior. Thus <defect, defect> is the sole equilibrium of the prisoner's dilemma.

Both the stag hunt and the prisoner's dilemma have many interpretations relevant to political and social philosophy. The stag hunt appears in Hume's *Treatise* in the form of a meadow-draining problem. Taking a modern example, James (2012) argues the stag hunt is representative of the situation we currently face with our global trade partners and hence is instructive when thinking of justice and collective-action on the international stage. For instance, countries morally motivated to adopt regulations that abate carbon emissions may fail to implement such policies without proper assurance others will follow suit (unilateral action is costly and does little to mitigate climate change). This situation is essentially a stag hunt; although all involved are eager to fulfill their moral duty and adopt pro-environment legislation, such activism is riskier than sticking with the status-quo.[9]

Skyrms, echoing Hume and Rousseau, contends the stag hunt aptly models the strategic situation underlying the formation of a social contract.[10] For the state of nature to be difficult to transcend, it must be a Nash equilibrium. This corresponds to the stable but inefficient <hare, hare> equilibrium. Opting to form a social contract is inherently risky but is beneficial to all if realized. Analogously, hunting stag is inherently risky but pays great dividends when successful.[11] Another line of literature, which can be traced to Rawls and Gauthier, takes the one-

[9] This is of course a complex example, and many pertinent details that would complicate the analysis are left out (for example, there may be disagreement among these morally motivated individuals as to what course of action is best. See Kavka (1995) for more about the dilemmas morally perfect people may find themselves in).

[10] This sentiment is shared by Edwin Curley in the introduction to *Leviathan* (1994). Other arguments in favor of the stag hunt can be found in Michael Moehler (2009). Additionally, although Kavka (1986) is often taken as arguing that the prisoner's dilemma is the best representation of Hobbes' state of nature, there is evidence that he believed the stag hunt was a more apt representation. On pages 113 to 116, Kavka introduces a new game he refers to as the N-person quasi-prisoner's dilemma and argues that if we are being very careful, it is technically *this* game that individuals in the state of nature play. Although it appears he failed to notice it, this *quasi-prisoner's dilemma is just the N-person stag hunt.*

[11] Binmore (1993) has a slightly different interpretation of the stag hunt game. He conceives of the stag hunt as capturing the strategic situation we face when determining whether to go through with reform or not. In this case the "state of nature" equilibrium is interpreted as the status quo, and the "social contract" equilibrium is interpreted as the new and improved state of affairs. Thus the results of this paper can be thought of as shedding light on both (i) the question of when a group can transcend the state of nature and (ii) the question of when reform is possible. For the

shot prisoner's dilemma to be the correct representation of the state of nature.[12] While it is compelling to view the prisoner's dilemma as a social contract game, Skyrms (2004, 4-6) demonstrates that when the strategy choice is between "grim trigger" and "always defect," deciding how to behave in a repeated prisoner's dilemma is strategically identical to that of deciding how to behave in a one-shot stag hunt.[13] Thus, there is a sense in which even if the interaction between agents in the state of nature is best modeled by the prisoner's dilemma, the repeated nature of these interactions directs our attention to the stag hunt.[14]

	Cooperate	Defect
Cooperate	R, R	S, T
Defect	T, S	P, P

Table 2.2: Normal form of the prisoner's dilemma with T > R > P > S.

A recent project among philosophers and social scientists seeks to determine under what circumstances stag hunters flourish.[15] Unfortunately, the replicator dynamics rarely leads to the stag hunting equilibrium. Additionally, it has been shown that a stochastic evolutionary system spends the majority of its time at the hare hunting equilibrium. Surprisingly, we'll see that

---

rest of the paper, we'll refer to the stag hunt game as a social contract formation game, although we believe Binmore's interpretation equally applies.

[12] In *A Theory of Justice* Rawls claims Hobbes's state of nature is the classical example of the prisoner's dilemma.

[13] Grim trigger commands agents to cooperate until their counterpart defects, in which case one plays defect for the remainder of the interaction. The grim trigger is not the only strategy that converts the repeated prisoner's dilemma to the one-shot stag hunt. For example, the popular tit-for-tat (imitate the act performed by your counterpart in the last round of the game) can also convert the repeated PD to a one-shot SH when the only other available option is always defect.

[14] One may wonder if it makes any sense to talk of repeated interactions in the Hobbesian state of nature (I thank an anonymous referee for bringing this criticism to my attention). If I cooperate while my counterpart defects, I am dead and the game screeches to a halt. Death may be one outcome, but I believe it is premature to assume that this is the only outcome or even the most likely outcome of confrontation in the state of nature. Kavka (1986), for example, suggests that one may attack or "anticipate" (i.e., defect) against another in the state of nature for the express purpose of physically dominating one's opponent and thereby forcing him to protect you against others. Vanderschraaf (2006) argues that it is implausible that one could master another in the state of nature for more than a brief time, instead interpreting the defect or "anticipate" strategy as the "seizure of some of their power. Often one anticipates against another by seizing some of the goods she possesses. One can also anticipate against another by occupying some of her physical and mental powers in a conflict." (247)

[15] Examples abound. For a few, see Alexander (2007), Wagner (2012), Smead and Huttegger (2011).

allowing agents to employ similarity-based strategies greatly increases the probability of arriving at the cooperative equilibrium.

2.3 Similarity-based Cooperation

Similarity-based cooperation is a deceptively simple concept. Agents condition their behavior in a game on a salient trait possessed by their counterpart. A trait could be anything from armpit scent to visual cues such as skin color, cultural emblems, or even hairdos. The agent cooperates if his counterpart's trait is sufficiently similar to his own. If the agents play the prisoner's dilemma, it is easy to see why similarity-based strategies will not result in sustained cooperation: a mutant possessing the correct trait defects and performs exceedingly well.

William Hamilton was the first in the biological literature to suggest the possibility of a similarity-based strategy.[16] Hamilton invites us to imagine a gene that regulates both the presence of an observable trait in the organism and the organism's propensity to cooperate with those possessing this trait. This so called "green-beard effect" is widely considered unfeasible.[17] While the possibility of hard-wired, similarity-based cooperation appears remote, theoretical biologists and social scientists have not abandoned the project. Many have attempted to show altruistic behavior can evolve with the help of similarity-based strategies.[18] For example, Riolo et al. posit a model with a continuum of potential traits. Individuals then determine whether to behave altruistically based on the Euclidean distance between themselves and their counterpart in trait-space.

One common finding is that cooperation in the prisoner's dilemma is possible but requires a cycling of traits. This so-called chromodynamics is due to the invasion of mutants. A group of cooperating agents all possessing green beards is invaded by a mutant green beard that defects when interacting with other green beards. Cooperation is still possible but requires that the agents condition their behavior on a new trait. Thus, green beards may be followed by purple beards which in turn are followed by blue beards, etc. The existence of cycling in nature seems

[16] Robson (1990) was the first to formally demonstrate how costless pre-game signals can foster high levels of cooperation in the prisoner's dilemma, albeit for brief periods of time.

[17] Richard Dawkins, in the *Extended Phenotype*, both coins the term "green beard" and argues against the possibility of such similarity-based strategies.

[18] For example see Antal et al. (2009), Axelrod et al. (2004) and Jansen and Baalen (2006).

doubtful because implicit in these theoretical models is the assumption that trait mutations occur much more rapidly than strategy mutations. In fact in some models it is assumed traits mutate on the order of two whole magnitudes faster than strategies mutate.[19] While it is questionable whether similarity-based strategies can sustain high levels of cooperation in the prisoner's dilemma, we'll see that it is much better suited to promote cooperative behavior in the stag hunt. It thus comes as little surprise that the strategic situation underlying one of the few documented cases of the green-beard effect in nature is reminiscent of the stag hunt.[20]

2.4 The Model

This simple model is loosely based on work in Riolo et al. (2001). Individuals are assigned two values, a trait (also referred to as a tag) and a tolerance level, both represented by numbers drawn from a discrete uniform distribution from zero to one with 0.001 intervals. If individual *i* has a tolerance level of Tol_i, he will behave in the following manner when interacting with a second individual, *j*:

Stag if $|Tag_i - Tag_j| < Tol_i$

Hare if $|Tag_i - Tag_j| \geq Tol_i$

A tolerance level of zero means one is an unconditional hare hunter, unwilling to cooperate even with an individual endowed with identical traits.[21] In Riolo et al. individuals with the same trait were artificially forced to cooperate with each other, and this strong assumption was in part responsible for their favorable results.[22]

---

[19] David Hales points this fact out. Additionally, Hales demonstrates through the use of simulation that cooperation is highly unlikely when traits and strategies mutate at the same rate.

[20] Specifically, the amoeba Dictyostelium discoideum possesses a "green beard" gene. When food is scarce the single-celled organisms huddle together and congeal into a mass perched upon a stem. If they are able to successfully coordinate, the majority of them survive; but if they fail to aggregate, the individual amoeba will perish. The gene csA encodes for the cell adhesion gp80 protein. Those without this protein are left behind when the aggregation process begins. The underlying strategic interaction here resembles a stag hunt. Establishing a protective sanctuary is possible but requires large-scale cooperation. Going it alone minimizes uncertainty but leads to a sub-optimal result. See Queller (2003) for more details.

[21] Note that an agent may also be willing to cooperate with those outside of the [0, 1] interval of trait-space.

[22] Sherratt and Roberts provide a detailed criticism of a few key assumptions in the Riolo et al. model. Relaxing these assumptions led to low levels of cooperation in the prisoner's dilemma, but as we will see, does not have the same effect when the stag hunt is considered.

We'll start with a population of 100 agents all with randomly sampled tags and tolerance levels. Each agent will be randomly paired with ten other individuals to play the stag hunt. Each interaction will result in a payoff.[23] We'll call the sum of these and only these ten payoffs the individual's "total payoff" (TP).[24] Agents will then be randomly paired once more for an imitation period. If an agent has a lower TP than her counterpart, she will adopt both the trait and tolerance level of the other player. If the two have the same TP, no imitation occurs.[25] We'll make slight alterations to the baseline model, and eventually introduce mutations to traits and tolerance levels.

It is worth being explicitly clear at this point that while many of the papers cited in section 3 pertain to biological evolution, the model explored in this paper is a model of cultural evolution. The agents in the model are presumed to be boundedly rational humans, striving to maximize their payoffs by imitating those more successful than themselves. While imitation makes perfect sense with respect to the publically advertised "tags", it is less clear how the imitation process works when it comes to tolerance levels. Tolerance levels are private traits, meaning one cannot directly observe the tolerance level of another. Yet one can attempt to infer the tolerance level of another by observing them over the course of many interactions. We modify the baseline model so the imitation of tolerance levels occurs with some noise– i.e., when imitating another, the tag is perfectly imitated but the tolerance level is imperfectly imitated. Incorporating noise into the imitation process does not appear to affect our results.[26] Yet as the size of the group grows, it is implausible to assume that each individual tracks the past behavior of each and every other group member. So how can individuals make an educated guess as to the tolerance level of their imitation partner? It is common in these game-theoretic models to assume the existence of some

[23] The payoffs we will use are S = 3, H =2. As the ratio of S to H increases the basin of attraction for the stag hunting equilibrium expands. For example, if we use the payoff S = 15, H = 7 our results greatly improve. These payoffs are taken from Brian Skyrms (2002).

[24] It may be the case the agent plays the stag hunt more than ten times, but only these ten payoffs will count toward her TP.

[25] Models of imitation are often formally equivalent to the more traditional replicator dynamic. See Jorgen Weibull (1995) for details.

[26] We alter the baseline model in the following way: whenever an agent imitates the tolerance level of another, her new tolerance level is the tolerance level of her imitation partner plus a draw from the Normal distribution centered at zero with a standard deviation of .01. Running these new simulations, we recover the results of section 5 for a variety of parameter settings.

sort of social mechanism that can provide agents with detailed information about others in the population. For example, individuals could attain a rough estimate of the tolerance level of another through the help of a gossip network, responsible for keeping track of others past behavior. Such networks are frequently assumed but not explicitly modeled in the evolution of cooperation literature.[27] Even a somewhat unreliable and noisy gossip network would nonetheless enable agents to acquire a rough estimate of the tolerance level of others in the population.

2.5 No Mutations

We begin by running 1,000 trials of the baseline model, described above. We find that without mutations, *all of these simulations result in universal stag hunting*—all random pairings result in both agents opting to hunt stag. What occurs is that the randomly distributed agents begin to cluster together in trait-space, making it easier for them to cooperate.[28] Clustering in trait-space is vital to establishing high levels of cooperation and is powerful enough to promote the formation of a social contract even when the agents themselves are not particularly tolerant.

Clustering occurs by chance. Two individuals relatively close to each other in trait-space happen to interact with one another. If they have sufficiently high tolerance levels, then they'll reap the benefits of stag hunting. These two agents typically outperform those with relatively higher levels of tolerance because the higher one's tolerance level, the more likely one is to be let down by one's partner—she'll go for hare while you opt for stag. Thus those with higher tolerance levels will imitate one of the paired agents with lower tolerance, adopting both their trait and tolerance level. The more populated this region of trait-space is, the more likely individuals from this region will interact with each other and successfully coordinate on stag hunting. This results in a positive feedback loop and before long the entire population is concentrated in this small interval of trait-space, all with sufficiently high levels of tolerance to cooperate with each other.

[27] See, for example, the large literature on indirect reciprocity. Some of the more famous papers which assume a gossip network in the background are Nowak and Sigmund (1998), Ohtsuki and Iwasa (2004) and Brandt and Sigmund (2004).

[28] This is the case even if we limit the initial tolerance levels by drawing instead from a more restricted discrete uniform distribution of zero to 0.1. By the end of the simulation, agents have very low levels of tolerance, but stag hunting is still possible because the population clusters up tightly in trait-space. This cluster is stable because unlike the prisoner's dilemma, once all are cooperating it no longer pays to have low tolerance.

This finding is robust, so much so that even including a substantial number of unconditional hare hunters (tolerance equal to zero) in the population does not prevent social contract formation. Table 2.3 shows that there are still a substantial number of simulations that converge to the cooperative equilibrium even in cases where half the population is initially an unconditional hare hunter.

Number unconditional hare hunters	0	10	20	30	40	50	60	70	80
Percentage of sims go to stag eq.	100	100	98	80	57	41	20	10	2

Table 2.3: Percentage of simulations that go to universal stag hunting as a function of the number of unconditional hare hunters in the initial population.

We now relax a key assumption that seems responsible for our pleasant results. We have seen that if agents cluster in trait-space, cooperation is almost inevitable. However, this assumes that agents can easily change their traits. Relaxing this assumption may reduce the prospects of wide-scale cooperation. What does it mean for a trait not to be easily imitated? As put forth in the opening sections, traits are just observable features upon which the players condition their hunting behavior. This is an extremely broad definition and encompasses a number of different features, such as clothing, cultural or religious symbols and even physiological attributes such as height, weight or race. Note that there is an important distinction between clothing and physiological features. Clothing, hairdos and shoes can all be manipulated much more easily than one's weight or race. We'd say race is a *sticky* trait while clothing is a *plastic* trait: in the short span of an afternoon one can easily alter one's wardrobe, but one can never change race. Style and physiological features can be thought of as extreme cases. Accents are an ideal example of an intermediary case: accents are malleable but cannot usually be altered in the short term.

We'd expect more inflexible traits to inhibit clustering and thus lead to low levels of cooperation. To assess the effect trait plasticity has on cooperation, we run identical simulations as before, except with probability p the agent will adopt the trait of her counterpart, conditional of course on her counterpart having a higher TP. Thus with p=0.5 there is a 50 percent chance an

agent will adopt the trait of her more successful imitation partner.[29] We run simulations with values of p = 0.2, 0.4, 0.6, 0.8 and 1.[30] See Figure 2.1.

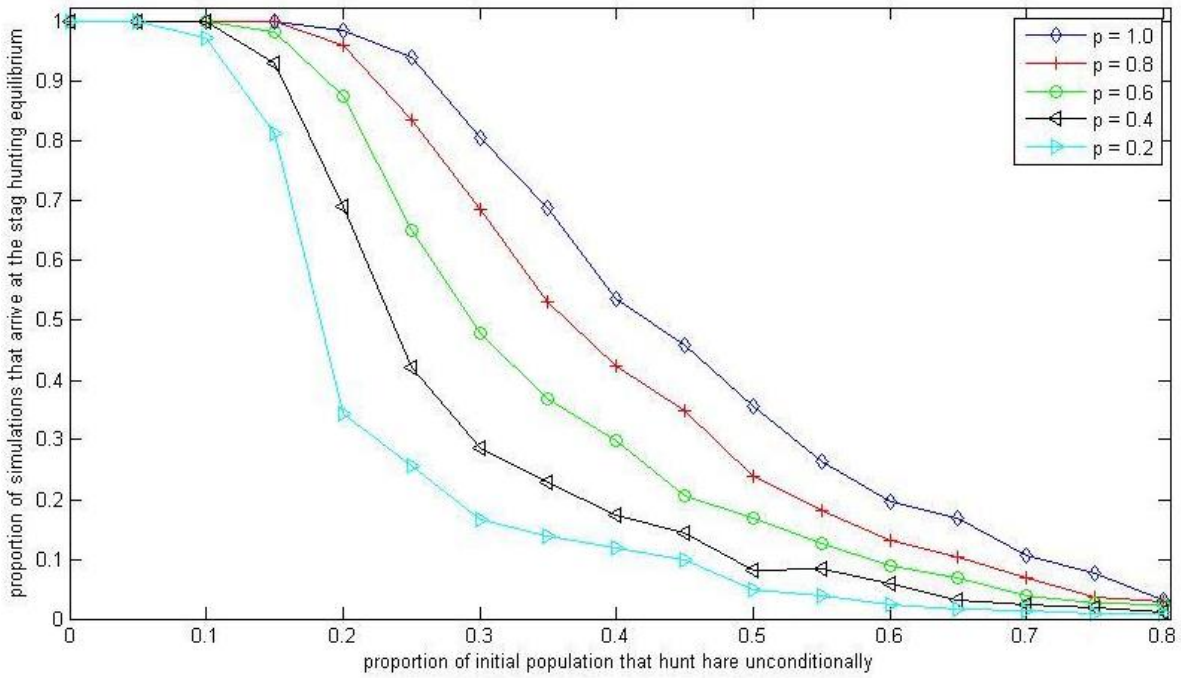

Figure 2.1: Percentage of simulations that arrive at the stag hunting equilibrium as a function of the number of unconditional hare hunters in the initial generation. Tag plasticity (p) varies from 0.2 to 1.

When there are no unconditional hare hunters in the initial population, trait plasticity is immaterial—all simulations result in universal stag hunting and sticky traits merely delay the inevitable convergence to the stag hunting equilibrium. If there are many unconditional hare hunters in the initial population, sticky traits result in low levels of cooperation.[31] Since traits are

[29] She will still always adopt the tolerance level of the imitation partner who outperformed her.

[30] Surprisingly, little research has been done on this topic. Axtell et al. (2006) investigate a bargaining model in which agents are endowed with permanent traits (p=0). Additionally, Skyrms and Zollman (2010) in this journal examine a model of agents with permanent traits (p = 0) playing the hawk-dove game. We are not aware of any work that investigates intermediary cases where p is between zero and one.

[31] When p = 0 (that is, traits are completely fixed and cannot be changed) and there are no unconditional hare hunters in the initial population, the population neither goes to universal stag hunting nor universal hare hunting. Instead individuals have moderate levels of tolerance, making them able to cooperate with those near them in trait-space, but not able to cooperate with all. Not surprisingly, as the number of unconditional hare hunters initially in the population

sticky, two cooperating individuals may not immediately attract others to their location in trait-space. In general clustering has to occur quickly because as more time passes it becomes more likely that the founding members of the cluster either imitate the tolerance or trait of an individual not in the cluster. Thus sticky traits make the early stages of cluster formation extremely fragile.

2.6 Mutations

The previous section demonstrated the essential role intolerance and homogeneity have in the evolution of cooperation. Somewhat intolerant individuals with similar traits cooperate and over time draw a majority of the population to their region in trait-space. This clustering means stag hunting is just about guaranteed. Overall, the social contract is possible but comes at the price of diversity and results in the proliferation of rather intolerant agents only willing to cooperate with a thin slice of trait-space. However this is not the end of the story. With reasonable mutation rates clustering can still occur, but the lack of diversity and tolerance is transient.[32] In the long run, cooperation is possible and diversity can be preserved. We can have our cake and eat it, too.

In the prisoner's dilemma, mutations cause problems. Mutants with lower than average tolerance exploit the members of the cooperative cluster. A new cluster emerges elsewhere in trait-space and establishes high levels of cooperation until mutants once again invade this new group. This cat and mouse game is what Riolo et al. observed. Such cyclic behavior is not present when we consider the stag hunt. Once the population clusters in trait-space, intolerant mutants will not thrive. They will unnecessarily refuse to cooperate with fellow group members willing to hunt stag. As long as the number of such mutants is low, they will be outperformed by those with high tolerance.[33]

---

increases, fewer and fewer of the interactions between individuals in the population consist in both hunting stag.

[32] Since our model is intended to explicitly model cultural evolution, we interpret mutations as either experimentation on the part of the agent (they make a deliberate choice to try a new strategy) or a failure on their part to correctly implement the strategy they choose (a trembling hand, as it is known in the economics literature).

[33] For our purposes we will model mutations in the following fashion: traits and tolerance levels are perturbed by a draw from a normal distribution with mean zero and a standard deviation of

The expanding circle

To get a better sense of how mutations affect long-term behavior, we'll first start by examining the case in which the agents have already successfully clumped together and formed a social contract. Consider the dynamics of such a group of agents clustered in trait-space. We start with 100 individuals all with the same trait (0.5) and miniscule tolerance level (0.01). Additionally, with a 10 percent chance, a given agent's tolerance will be perturbed by a pull from N(0, 0.1). An agent's tag will for now not be perturbed. What occurs is the following: mutants with a tolerance level lower than the initial population average fare poorly. Due to their ultra-low tolerance level, they foolishly hunt hare with a community of individuals all willing to hunt stag. Unsurprisingly, mutants with a tolerance level equal to or greater than the initial 0.01 continue to successfully coordinate on stag hunting. *Thus there is a weak selection for higher tolerance levels*, so much so that after 500 generations the average tolerance level of the population has risen to an astounding 0.624.

We will now include mutations to traits to the above framework. We find that for a number of parameters this does not prevent the population from continuing to cooperate. Tolerance increases just as it did in the absence of trait mutations. What additionally occurs is a diffusion of agents throughout the trait-space. When the average tolerance of the cooperative cluster is low, any trait mutations will likely be selected against—mutants are too different to cooperate with. When the average tolerance of the population increases, mutants can thrive. Diversity is slowly regained as average tolerance increases.

0.1. We allow tolerance levels to go below zero and restrict our tags to the interval [0, 1]. Riolo et al. modeled mutations in a similar fashion, except tags were replaced by a draw from the uniform distribution U[0,1]. This alteration will not significantly change the qualitative results of our simulation.

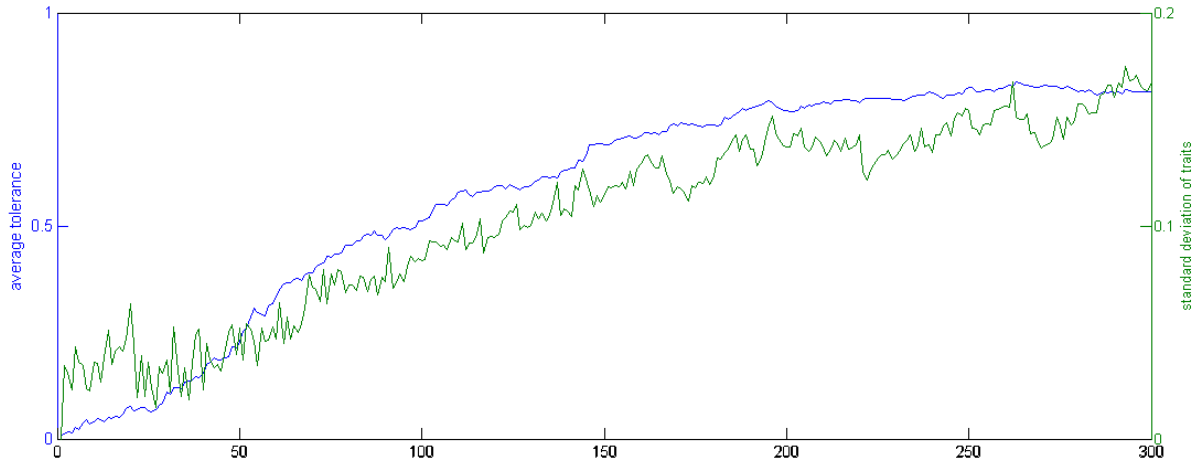

Figure 2.2: Average tolerance (blue) and standard deviation of the distribution of traits (green) over the course of the first 300 generations of a simulation. This data comes from one simulation consisting of 100 individuals all with an initial tolerance level of 0.001 and a trait values of 0.5.

This process is slow. Diversity is slowly regained because (i) perturbations to traits are small and (ii) traits that stray too far from the cooperative cluster fare poorly because they fall outside of the cluster's tolerance radius. Each generation has a slightly higher average tolerance level than previous generations, expanding the "radius" of cooperation. This circle expands slowly until agents from both extremes of trait-space can peacefully cooperate. In other words, with enough time, a highly intolerant and homogenous population can be transformed into a diverse and tolerant one.

When traits are sticky stag hunting can still thrive, but a lower trait mutation rate is necessary. If the trait mutation rate is high, then many agents accumulate on the periphery of the cooperative cluster, making it easier for those with a higher than average tolerance level to be let down by their counterparts, thus allowing intolerant agents to flourish. This then pushes the population to the hare hunting equilibrium. If the trait mutation rate is low, cooperation is possible and in the long run the familiar dispersion of agents throughout trait-space occurs.

Hence, sustaining a social contract is possible when traits are sticky, but hinges on the trait mutation rate.[34]

The contracting and then expanding circle

Let's put all of this together. We now start with individuals spread out randomly in trait-space and allow mutations to both tags and tolerance. As in Section V, we see a clustering in trait-space and all within the cluster are successfully hunting stag. Once this occurs we then observe a steady increase in tolerance accompanied by a slow spread of agents in trait-space. This two-stage process demonstrates that a successful social contract requires a certain amount of cohesion; however, this structure need not be permanent (See Figure 2.3). Once a social contract emerges, more and more of the trait-space can slowly participate.

The contracting-expanding dynamic is visually striking if we consider two dimensional trait-space. Agents still possess a tolerance level but now have two traits as opposed to just one. The distance between two agents is simply the Euclidean distance between points in two dimensional space. (See Figure 2.4).

Once again our results are less hopeful if traits are sticky. We run a hundred simulations in which traits are imitated with a 0.15 chance ($p = 0.15$) and find only 59 percent of these simulations result in universal stag hunting. As it becomes increasingly difficult to imitate the traits of others, we are less likely to observe this contracting-expanding dynamic. The percentage of simulations that underwent the contracting-expanding dynamic are 2, 13, 59, 76, 87, 91 and 100, for a p of 0.05, 0.1, 0.15, 0.2, 0.25, 0.3 and 0.4, respectively.[35] Sticky traits affect both stages of the dynamic. Low trait plasticity makes it more difficult for clustering to

[34] The less sticky a trait is, the less trait mutation rates matter. When traits are completely plastic we stay at the stag hunting equilibrium when the trait mutation rate is low (0.01) as well as high (0.15). Of course, if trait mutations are too frequent (0.5) then clustering is no longer possible and we revert to the hare hunting equilibrium.

[35] In these simulations there was a 10 percent chance for both trait and tolerance mutations, mutations for both traits and tolerance were sampled from N(0, 0.05) and the initial tolerance levels were determined by the uniform distribution from zero to one-half.

occur. Additionally, if clustering is successful, too many mutants on the periphery of the cluster can destabilize the group and result in the population settling on the hare hunting equilibrium.[36]

2.7 Discussion

This paper investigates the deep connection between diversity, tolerance and the social contract. We found, in particular, that similarity-based strategies can promote cooperative behavior in many scenarios and, surprisingly, homogeneity and intolerance are often essential intermediate steps toward the formation of a social contract. However, as vital as homogeneity and intolerance are to cooperation, both are not long-term features of the population. Once the agents are cooperating, tolerant mutants will prosper and soon all of trait-space can participate in a thriving social contract. In the long run, a social contract amazingly does not come at the price of diversity.

These fortuitous results suggest a natural moral progression, namely, that intolerance and homogeneity often pave the way to diversity and wide-spread cooperation. Such a transition has been noted in modern times by Peter Singer. Singer observes a natural tendency for the circle of altruism to "broaden from the family and tribe to the nation and race" and eventually go so far as to "extend to all human beings."[37] Figure 2.3 and Figure 2.4 nicely illustrate such a movement. Over time a social contract transforms from an exclusive enterprise to one that allows all to participate.

This is all with one wrinkle though. Singer takes this progression to be the result of reason, going as far to suggest the expanding circle is either an "accident of history […] or the direction in which our capacity to reason leads us." (113) This appears to be a false dichotomy. The results uncovered in this paper are neither accidental nor are they driven by deliberate moral thought and reason. Instead, the border of the circle expands due to simple agents merely

[36] However, if we reduce the frequency of trait mutations, our results change for the better. Nearly all simulations arrive at the stag hunting equilibrium when the frequency of trait mutations is one-tenth that of the frequency of tolerance mutations. It seems sensible to assume that since traits themselves are sticky, mutations to them occur relatively infrequently.

[37] Peter Singer, *The Expanding Circle*, page 120. While we and Singer seem to slightly differ in our use of the word cooperation (Singer is concerned with altruism – costly acts which help others – while our concern has been primarily with collective action) it is immensely striking that the general 'moral' we both come to remains the same. Namely, there is a tendency for individuals to naturally cooperate with less and less restrictive subsets of the population.

imitating those who are successful. The continual stream of mutations and experimentation is enough to cause boundedly rational individuals to become more tolerant. No moral reflection or reasoning is necessary.

Thus our results seem more in line with Milton Friedman than Singer. Once a social contract is established, it is in the agent's interest to have a higher tolerance level. Yet there is an important difference that distinguishes our account from Friedman's. On Friedman's account, one always has self-interested reasons to be more tolerant. This is not true in our model. Too much change too quickly will be selected against. Consider the situation in which the population is crammed into a region of trait space and for the most part all have the minimal levels of tolerance that guarantee near universal stag hunting. If an agent increases her tolerance too much, then she'll cooperate with those outside the cooperative circle unwilling to cooperate with her. Ergo she would have benefited from being less tolerant. Increasing one's tolerance level by small increments is individually rational, but becoming an unconditional stag hunter may not be. As a result, the transition from an intolerant society to a tolerant one is a slow and gradual process – too much change in a short time period will be selected against.[38]

It is also worth mentioning that so far we've been implicitly assuming that there is just one viable social contract. This of course is not the case. In reality, there are a plethora of contracts and there exist real differences between the various feasible social contracts. Some, for example, are more efficient than others. How can we transcend an inefficient social contract? How is reform possible? In many ways, the problem of social contract *reform* is very similar to the problem of social contract *formation*. As a result, the solution we propose to the problem of reform will be very similar to the findings we've uncovered in this paper. Traits and similarity-based strategies help facilitate institutional change. Consider a slightly modified stag hunt game with not two, but three equilibrium. The game is almost identical to the stag hunt game, except we allow for a third strategy.[39]

---

[38] On Friedman's account change occurs slowly as well, but for different reasons. Change is slow because it takes time for discriminatory firms to be weeded out of the market. On our account, change is slow because becoming *too* tolerant can be disadvantageous to the individual. At any given time period, there is an upper bound on how tolerant one could profitably be.

[39] We are assuming that social contracts differ just in how efficient they are. Of course some social contracts could be more equitable than others. Whether traits and similarity-based strategies can

	A	B	C
A	4, 4	0, 3	0, 2
B	3, 0	3, 3	0, 2
C	2, 0	2, 0	2, 2

Table 2.4: A modified stag hunt game.

Consider that initially individuals only had strategies B and C available to them. In other words, they were playing the game from Table 2.4 instead of Table 2.1. We will continue to assume that agents possess a trait and a tolerance level, and play B with those in their tolerance radius and C with those outside the radius. By the end of these simulations we'd end, as we did in Section 5, with all individuals spread out in trait-space, and all with a sufficiently high tolerance level to play B with any other member in the population. Now once this (B, B) social contract is established, let's introduce strategy A (perhaps strategy A was not a viable strategy before). How do we now make the transition from the (B, B) equilibrium to the (A, A) equilibrium? Consider the following series of events. By chance two agents very close to each other in trait space experiment and each adopt a strategy which allows them to play A against those close to them in trait space and B against anyone outside of this "A tolerance radius." Just as in Section 5, these two individuals will do slightly better than the rest of the population. This will in turn attract more of the population to this area of trait space, meaning even more agents will be willing to play A, as opposed to B. In the end, we will recover the familiar contraction-expansion dynamic identified in Section 6 – the final result being a transition from the (B, B) equilibrium to the (A, A) equilibrium.

In this way, traits allow those yearning for change to successfully interact with like-minded individuals while simultaneously ensuring reformers do not waste their time playing A with those unwilling to deviate from the status quo. This is fascinating for it demonstrates there is a natural and straightforward way in which reform is possible. Even more surprising, *this explanation does not rely on group selection*. Binmore (2005), for example, has argued informally that group selection is responsible for ensuring that efficient social contracts are adopted by societies.[40] We should expect evolution to "succeed in selecting one of the efficient equilibrium [because] societies

effectively facilitate transitions from inefficient to efficient social contracts when we take fairness and equity into account will not be explored in this paper.

[40] For an interesting assessment of this thesis see Huttegger and Smead (2011).

operating [with an efficient social contract] will grow faster." If we assume "societies cope with population growth by splitting off colonies which inherit the social contract of the parent society," efficient social contracts should abound. Traits and similarity-based strategies can successfully allow for the transition from inefficient social contracts to efficient ones. Group selection may speed up this process, but it is by no means required.

Finally, while the result of this paper are promising, we should remember the rather intuitive finding that the social contract is less likely when traits are sticky. In the presence of persistent diversity, establishing an efficient social contract seems improbable. Further exploration of sticky traits seems fruitful. For instance, when individuals have multiple observable traits can agents learn to only take into account those traits that are plastic? If so, cooperation seems unavoidable. Imagine a population of agents each with two traits, one plastic and the other permanent. Individuals may measure their distance in trait-space by calculating the Euclidean distance between them in two dimensional space. However, agents could just as easily only base their decision on how closely they resemble their counterpart with respect to one trait. For example, individuals may decide to base their behavior solely on how closely they resemble their counterpart with respect to the plastic trait. How would a population of racists (those conditioning their behavior on fixed traits alone) and open-minded individuals (those conditioning their behavior on plastic traits alone) fare? This is an open question.

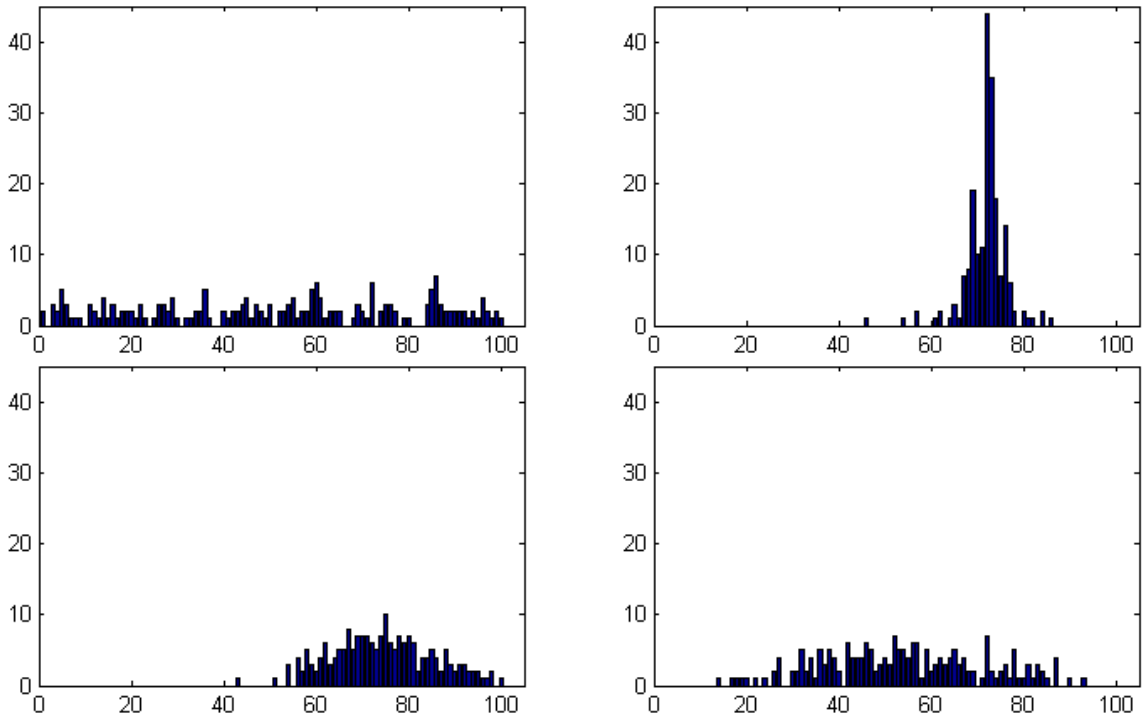

Figure 2.3: Distribution of traits. T = 1 (top left), T = 150 (top right), T = 1000 (bottom left), T = 2000 (bottom right). Initial tolerance was drawn from the distribution U[0, 0.1] and p = 1.

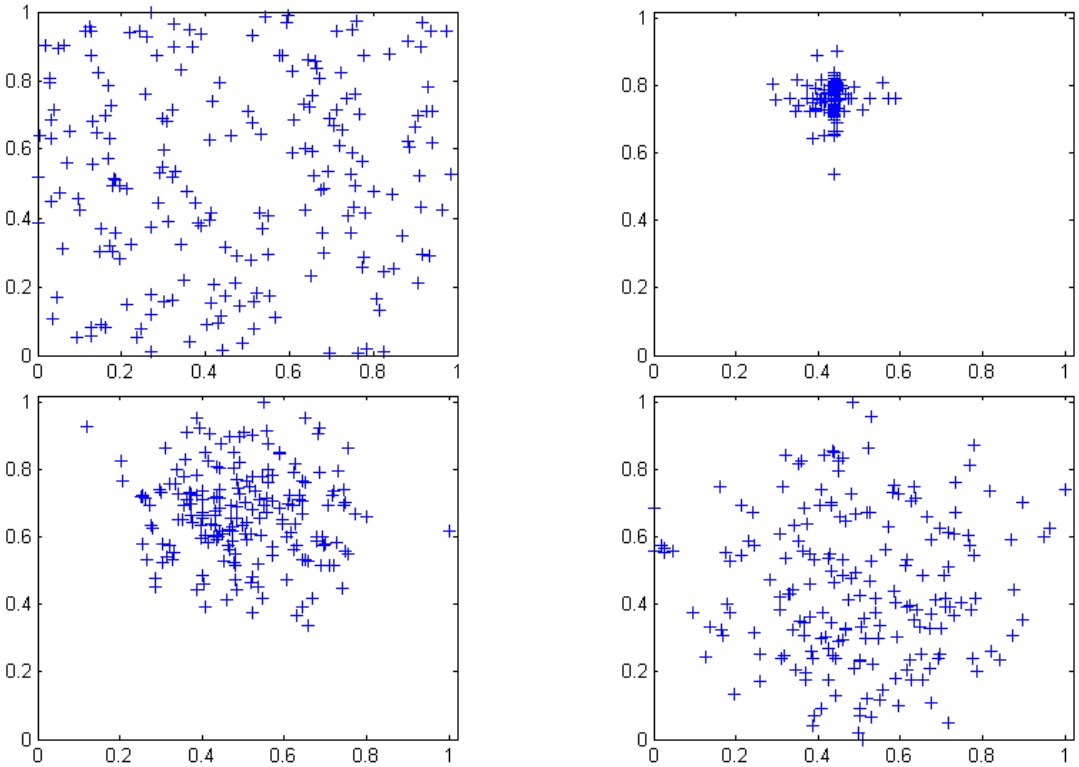

Figure 2.4: Two trait case. Distribution of traits in 2-dimensional trait-space. T = 1 (top left), T = 40 (top right), T = 1000 (bottom left), T = 2000 (bottom right). Initial tolerance was drawn from distribution U[0, 0.1] and $p = 1$.

Chapter 3

RACISTS AND MINORITIES IN POPULATION GAMES

3.1 Introduction

Racism is an all too common phenomena. From outright violence to preferential treatment, racism still plays a significant role in modern society. Unfortunately, while philosophers have recently done much to investigate the concepts of race and racism, little progress has been made to better understand the dynamics of racist behavior. This latter project is especially vital if we aim to understand the conditions under which discriminatory norms flourish and decline.

We champion a game-theoretic approach that allows us to precisely study the dynamics of racist behavior. We model race as an immutable, permanent and costless pre-game signal and racism as a propensity to condition one's behavior on the immutable, fixed signals of others. With this basic representation of race and racism, we turn our attention to the study of bargaining behavior (the Nash demand game) and collective action (the stag hunt game) in a population of diverse agents. Evolutionary game theory and computer simulations allow us to paint a detailed picture of how these social contracts unfold in the presence of racist behavior.

Importantly, our approach enables us to study how racial minorities fare when in the presence of race-based strategies. This theoretical investigation is especially significant because minority groups are often the target of discriminatory norms. Our game-theoretic investigation reveals that minorities in certain circumstances are much more likely to be the victim of discriminatory norms. For example, we find that in the Nash demand game minority status translates into a weak bargaining position.

Lastly, we investigate one possible means of reducing racist behavior. Past theoretical studies of discrimination have been unable to study racism reduction because these studies often hold fixed individual behavior. For example, Schelling's famous segregation model assumes all individuals have a weak preference to surround themselves with those of their own race. These preferences are *static* and are not allowed to evolve. More recent investigations, for example Hammond and Axelrod (2006), allow for more freedom but are primarily interested in providing

an evolutionary account of ethnocentric behavior. Our project departs from these past studies because our concern is to explore a means of reducing ethnocentric or racist behavior.[41] We find one means of promoting high levels of cooperation and reducing racist strategies is to allow individuals to condition their behavior on the plastic, or malleable, signals of others.

Our paper will proceed in the following fashion. In section 2, we operationalize both race and racism and demonstrate how these concepts can be captured in our game-theoretic framework. In Section 3, we investigate how minorities fare when playing the Nash demand game and in Section 4 we discuss behavior in the battle of the sexes game. Section 5 centers on the stag hunt game and it is in this section we explore how discriminatory norms can be jettisoned. Section 6 concludes and connects our results with prior work in social philosophy, social psychology and biology.

3.2 Racism and Cheap Talk

Recent philosophical work on race is replete with conceptual analysis. Mallon (2006, 2007) provides a thorough overview of recent work on the concept of race. For our purposes we take a page from historian George Frederickson and utilize his intuitive notion of racism. Frederickson (2002) notes that throughout history ethnocentrism is often conflated with racism. While these two behaviors may prime facie seem indistinguishable, there are important differences between race and culture that need to be taken into account. While racism can be thought of as differential treatment based on race, Frederickson conceives of ethnocentrism as differential treatment based on cultural or religious identity. Thus one may avoid ethnocentric discrimination through the process of assimilation, i.e., changing one's religion or adopting the majority culture. Racial discrimination, on the other hand, is harder to circumvent due to the permanent nature of race.

For our purposes race will refer to an attribute of the agent which satisfies the following three features:

(1) The attribute is visible

(2) The attribute is immutable (i.e., the expression of the attribute cannot be suppressed)

[41] One notable exception is Grim et al. (2005) and Grim et al. (2008), to be briefly discussed in section VI.

(3) The attribute cannot be modified

We can capture the above three criteria by representing race as a fixed costless pre-game signal sent by the agent before engaging in a strategic interaction. The signal sent is immutable, i.e., agents cannot help but send the signal. The signal is also permanent – individuals cannot alter the signal they send.[42]

In addition to endowing agents with a fixed costless pre-game signal, we also allow individuals to condition their behavior on the signal of their counterpart. Thus one individual can now treat agents broadcasting different pre-game signals differently. An example of such a strategy in the prisoner's dilemma would be the following: cooperate with *Blue* individuals and defect when paired with *Red* individuals. In this situation there are two types of fixed signals, *Red* and *Blue*, and an agent's strategy specifies how to act when faced with either sort of individual.

While conventional wisdom predicts cheap talk has little to no effect on games that aren't purely cooperative, prior work on signaling has demonstrated that in an evolutionary context, cheap talk can have dramatic effects on behavior. Robson (1990) demonstrates how costless signals can sustain cooperation in even the prisoner's dilemma, albeit for brief periods of time. Signals can promote cooperation in other contexts as well. Skyrms (2004), for example, finds the introduction of cheap talk greatly increases the likelihood of avoiding the sub-optimal hare equilibrium in the stag hunt game.

Our approach is similar to that of Skyrms (2002, 2004). We study a population of agents randomly paired to play a game. The proportion of strategies in the population change as determined by the replicator dynamics, a standard model of biological evolution (Taylor and Jonker, 1979). We utilize the replicator dynamic because many forms of cultural learning and imitation are formally equivalent to this dynamic (Sandholm, 2010). In the sections that follow we investigate a diverse population of agents tasked to play a variety of different games.

[42] If signals instead were to represent cultural or religious affiliations, signals would be mutable and malleable.

3.3 Minorities in the Nash Demand Game

We now turn our attention to bargaining games, specifically, the Nash demand game. Bargaining games have recently acted as a focal point for those interested in naturalized ethics and the evolution of moral norms. Alexander and Skyrms (1999) study bargaining behavior on a network and contend the Nash demand game can nicely capture two central concerns of distributive justice: equality and efficiency. Additionally, Ken Binmore (1994, 1998, 2006) places bargaining games at the center of his impressive work on justice. The game of life, Binmore argues, is filled with a number of varied strategic situations, but if one were forced to reduce the game of life to a single strategic interaction one would be wise to select a bargaining game.

Although there are a plethora of formal models of bargaining, we restrict our attention to the more manageable Nash demand game. In this game two agents must divide a resource. The two players simultaneously demand a fraction of the resource and receive what they demand if the combined claims are equal to or less than the total value of the resource. Any remaining resource left after the bargaining process is destroyed. If the combined claims are greater than the total value of the resource, both individuals walk away empty-handed. We restrict our attention to a simple version of the game, in which 10 units of resource are available and players utilize one of three strategies: Demand 4 units, Demand 5 units and Demand 6 units (see Figure 3.1 below).

	Demand 4	Demand 5	Demand 6
Demand 4	4, 4	4, 5	4, 6
Demand 5	5, 4	5, 5	0, 0
Demand 6	6, 4	0, 0	0, 0

Table 3.1: Nash demand game with ten possible units of resource and three available strategies.

It is clear the three pure Nash equilibrium of this game are (4, 6), (6, 4) and (5, 5). There also exists a mixed equilibrium in which players demand four units two-thirds of the time and demand six units one-third of the time.[43] Which of these possible equilibria should we expect to converge to? Rational choice theory cannot provide a definite answer. However, evolutionary

[43] There also exists a completely mixed Nash equilibrium in which individuals mix over all three strategies.

game theory can give us a clearer sense of how play will unfold in the Nash demand game. Instead of just two agents playing the above game, we envision a population of agents randomly paired to play the Nash demand game. Individuals then adopt the strategy of those who were more successful than themselves. Under the replicator dynamics, a standard model of both cultural and biological evolution, the population goes to one of two stable states. In the fair division equilibrium all individuals play the *Demand 5* strategy. There additionally exists a polymorphic equilibrium in which two-thirds of the population plays *Demand 6* while the remaining one-third play *Demand 4*.[44] Although both of these arrangements are stable, the basin of attraction for the fair division equilibrium is much larger than the basin of attraction for the inefficient polymorphic equilibrium, meaning a fair division is much more likely to emerge. Do we arrive at such pleasant results if we allow individuals to condition their behavior on the race of their counterpart?

Axtell et al. (2000) investigate a slightly more complicated variant of the Nash demand game to address this question.[45] In their version of the game, individuals are endowed with one of two permanent signals, call them *Red* and *Blue*. Individuals are once again randomly paired to play the Nash demand game, but importantly, are allowed to condition their behavior on the signal of their randomly selected partner. Thus agents no longer have simple strategies such as Demand 4, but instead have conditional strategies such as "*if counterpart Blue, then Demand 4; if counterpart Red, then Demand 6.*" If we consider the signal the agents send as part of their strategy, a strategy can now be thought of as a 3-tuple. The strategy <R, 4, 6>, for example, instructs the agent to send signal Red, demand 4 when dealing with fellow Reds and demand 6 when interacting with Blues.[46]

Axtell uncovers that there are many new equilibria in this modified game. For example, there exists a new asymmetric equilibrium where *Red* agents demand six when interacting with *Blue* agents, and *Blue* agents in turn only demand four when interacting with *Red* agents. Additionally, when interacting with their fellow *Blues*, *Blue* agents demand 5. Likewise, when

[44] This polymorphic equilibrium is inefficient because when Demand 4 types meet each other they leave two units of the good on the table, while when Demand 6 types meet each other bargaining breaks down and both walk away empty-handed.

[45] Zollman and Skyrms (2010) engage in a similar exploration except with the hawk-dove game.

[46] There a total of 18 such strategies.

interacting with *Reds*, *Red* individuals demand 5.[47] In other words, a fair division exists within the *Red* and *Blue* populations, but interactions between *Red* and *Blue* individuals are significantly one-sided, with the *Reds* taking the bulk of the resource. Recall that individuals are incapable of changing the signal they send, making it natural to interpret these signals as permanent features of the individual, such as race.[48] Seen in this new light, Axtell's results demonstrate that permitting individuals to condition their behavior on race in a bargaining game can often result in the natural emergence of discriminatory norms.[49]

While Axtell has demonstrated discriminatory norms can naturally arise, he assumes the *Blue* and *Red* population are of equal size. This of course needn't be the case. Since minority groups have historically been the target of discriminatory norms, we adjust the relative size of the *Blue* and *Red* populations to determine if minorities are at a distinct disadvantage. Empirical studies strongly suggest minority status translates into dismal economic prospects and unequal access to vital resources such as health care and education. While the unequal treatment of minorities has been a well-documented phenomenon in a number of cultures and countries, this paper will provide a game-theoretic explanation of these glaring inequalities.

We first investigate the baseline case where the "races" – *Blue* and *Red* – are of equal size.[50] In the baseline case, slightly over 56 percent of the time the population is pushed to the equal split equilibrium, meaning the resource is split evenly between individuals of different races. The remaining cases settle at an asymmetric equilibrium, where one population demands 6 and the other population acquiesces and only demands 4. This is identical to Axtell's main

[47] This is not the only new equilibrium. For example, while there could exist an equal split between Reds and Blues, and there could in addition be a 6-4 split *within* the Red population, meaning two-thirds of the red population demand 4 of their fellow Reds and the remaining one-third demand 6 of Reds.

[48] Note that these results are only possible because signals are fixed. If individuals were instead allowed to change whether they send Red or Blue, the above arrangement would no longer be stable and all would quickly adopt the Red signal.

[49] Axtell and his co-authors did provide an explicitly racial interpretation and simply argued that their results demonstrate how the natural emergence of class hierarchy is possible.

[50] This is done using the two-population discrete-time replicator dynamics. Individuals in each racial population are randomly paired to play a game with either a member of their population or the other population. The composition of strategies in each population are governed by the following equation: $x_i(t+1) = x_i(t)\frac{F_i}{F}$, where $x_i(t)$ refers to the proportion of agents in the population who employ strategy i at time t, F_i is the average fitness of strategy i and F is the average fitness of the population.

finding. We now alter the relative size of the Blue and Red populations. This is done by changing the probability with which agents are paired to interact with Red or Blue individuals. Results are presented below in Table 3.2.

As the Red population shrinks in size the dynamics are less likely to lead to an equal division between the two populations. Furthermore, it is much more probable the population will settle on an equilibrium in which the minority Reds demand only 4 when interacting with the majority Blues. This tendency becomes more pronounced as the minority makes up a smaller fraction of the overall population. When the minority is one-tenth of the total population, a little over half of simulations result in the Majority Blues demanding 6 of the minority Reds.

Proportion of Reds	*5-5 split btwn. Groups*	*6-4 split favors Blues (majority)*	*6-4 split favors Reds (minority)*	*5-5 split within Blues (majority)*	*5-5 split within Reds (minority)*
0.5	56.43%	21.84	21.72	81.14	81.54
0.4	55.62	28.19	16.18	82.91	79.79
0.3	52.83	34.39	12.78	83.73	78.91
0.2	47.60	42.94	9.46	85.33	76.42
0.1	39.05	53.56	7.38	86.08	74.11
0.05	34.83	57.95	7.21	86.41	73.17
0.01	31.67	62.15	6.15	87.13	72.31

Table 3.2: results of simulations of the two-population discrete-time replicator dynamics. Each entry summarizes the results of 10,000 simulations.

When only a miniscule 1 percent of the total population consists of Reds, 62.2 percent of simulations result in the majority demanding 6 of the minority, while a staggering 6.2 percent of simulations result in the minority demanding 6 of the majority. When it comes to negotiating a division, minorities are at a considerable disadvantage, and this disadvantage becomes more pronounced as the minority shrinks in size. In the Nash demand game, *minority status translates into a weak bargaining position.*

The intuition behind this result is as follows. Recall the conditional strategy individuals possess. A strategy in this game consists of two "contingency plans." One plan guides the agent when interacting with Reds, while the other contingency plan guides the agent when interacting with Blues. If the Reds are the minority, then on average both Blue and Red individuals will interact with many more Blues than Reds. As a result, the contingency plan that guides

individuals when interacting with Blues changes rapidly, while the contingency plan responsible for instructing agents when interacting with Reds remains, for the most part, stagnant. Hence, both Red and Blue individuals will quickly adjust their "blue contingency plan," while the "red contingency plan" will be updated relatively slowly. In other words, the minority Reds will quickly adjust so as to best respond to the relatively inert (and randomly assigned) "red contingency plan" of the Blues. It is now clear why the minority so frequently ends up demanding the low amount of four from the majority – the strategy "demand four" is typically the Red's best response to the initial random distribution of Blues.

The difference in size between the two populations leads to a difference in inertia, which in turn results in the emergence of a discriminatory norm that favors the majority. Since the majority has more inertia than the minority, the majority continues to play their initially assigned "red contingency plan" long after the Red population has begun to alter how it interacts with the Blue population. This asymmetry of inertia, where the majority slowly adjusts its play while the minority readily adapts to the behavior of the majority, is the primary reason the minority is often left with the smallest slice of the communal pie.

Similar results on the role of inertia can be found in the game theoretic literature. Consider Peyton Young's study of a repeated bargaining game. Young (1992, 1996) explores a bargaining model where each individual has a finite memory of past interactions. An agent's behavior in the current round is to simply play whatever strategy is a best response to a random sample of past play, as recorded in her finite memory database. Young demonstrates that those with better memories are more likely get the upper hand in a bargaining interaction. This once again boils down to inertia – the longer one's memory, the more difficult it is for random perturbations to change what her best response is.

3.4 Minority Advantage?

We've shown above that minorities face a distinct disadvantage when bargaining. Yet bargaining games constitute just one small slice of the total number of strategic situations individuals may be confronted with on a daily basis. Does this minority disadvantage spill over to other settings? We see that it need not, and moreover there are circumstances in which it actually pays to be in the minority.

Consider a coordination game such as the battle of the sexes (BoS). In this game two individuals must arrange a meeting place. In our simple toy model we'll assume that agents can either meet at the coffeehouse or the teahouse. If both agents decide to head to the same establishment they are able to find each other. If they instead patronize different establishments they miss each other, and this is the worst outcome of the game. There is an important asymmetry, however. While it is assumed that both agents prefer meeting over not meeting, the agents have different preferences over where to meet. In particular, as Table 3.3 indicates, row player ranks meeting at the coffeehouse above a meeting at the teahouse, while column player has the opposite preference.

		Green (column)	
Brown (row)		Coffeehouse	Teahouse
	Coffeehouse	$\alpha + \varepsilon, \alpha$	$0, 0$
	Teahouse	$0, 0$	$\alpha, \alpha + \varepsilon$

Table 3.3: Battle of the sexes with options of Coffeeshop or Teashop. The parameters α and ε are both assumed to be strictly positive.

Now consider two groups, the Brown and the Green. Individuals from the Brown population have a visible marker which distinguishes them from their peers from the Green population. Now also consider the fact that those from the Brown population have an overwhelming preference for coffee, while those in the Green population almost unanimously favor tea (and especially *green* tea!). Just as in the previous section, individuals possess visible attributes (or markers) and thus can condition their behavior on the marker of their counterpart. Not surprisingly, a number of equilibria are possible. For example, everyone can ignore the marker of their counterpart and simply meet at the Coffeehouse. Likewise, Brown individuals can concede and meet Green counterparts at the Teahouse, while patronizing the Coffeehouse when interacting with their own. Each group desires to meet at their favorite establishment. Which group wins out in the end?

The answer depends on the relative size of the Brown and Green groups. If the groups are of the same size then the coffeehouse and teahouse both have the same chance of becoming the go-to hang-out spot when a pair of Brown and Green individuals attempt to meet.

Surprisingly, if we vary the relative sizes of the population we uncover that the minority group has an advantage! When Browns only constitute a small 5 percent of the overall population over 64 percent of simulations result in an equilibrium in which Browns and Greens meet at the coffeehouse – the preferred location of the Browns.[51] Once again inertia can shed light on the phenomena. The majority is very unresponsive to the behavior of the minority. This once again means that the minority adapts rapidly to the initial distribution of strategies in the majority population. In the Nash demand game that meant the minority more often than not demanded the low amount of four. For our Brown minority group, the best response to a randomly selected initial Green population is more often than not to go to the coffeehouse.

In light of these results it is apparent that we cannot say with surety that the minority is always at some sort of disadvantage. In certain classes of games, such as the bargaining games studied above, the minority is more likely to be disadvantaged, stuck demanding a low amount of the resource. In other circumstances, such as a BoS, the minority does well, in the sense that the population frequently goes to the equilibrium they favor. It is not the case that in all strategic situations the minority is at a disadvantage. However we can, as we have done in this paper, restrict our focus to particular classes of games and discover whether they confer an advantage to the minority. In the BoS of Table 3.3, for example, as long as epsilon is strictly positive it will always be slightly more likely that the population will settle on the equilibrium that favors the minority than one that favors the majority.

3.5 The Stag Hunt and Racism Reduction

In this section, we explore a means of reducing racist behavior in the context of the stag hunt game. Specifically, we examine the realistic setting where individuals possess both fixed and plastic signals and are allowed to condition their behavior on either the fixed or plastic signal of their counterpart. We demonstrate that allowing individuals to condition their behavior on either of these signals results in both a drastic increase in the amount of cooperation as well as a sizable reduction in the number of race-based strategies in the population.

[51] This is once again using the two-population discrete-time replicator dynamics, with $\varepsilon = \alpha = 2$. Interestingly enough the minority had a 30 percent chance of landing up at their sub-optimal teahouse equilibrium, while the majority had only a 25 percent chance of doing so. So there may be some dividends to being in the majority in this case.

	Stag	Hare
Stag	3, 3	0, 2
Hare	2, 0	2, 2

Table 3.4: Strategic form of the stag hunt game.

The stag hunt game is a simultaneous move game between two individuals. If both opt for stag, they are rewarded with a large payoff. If either instead hunts for hare, the hare hunter receives a small payoff, while the lone stag hunter receives nothing. (Stag, Stag) and (Hare, Hare) are both equilibrium of the game, and both have their merits.[52] Hunting for stag pays great dividends when successful but is inherently risky; hunting for hare yields a certain payoff but this payoff is mediocre. The stag hunt game is particularly interesting because, like the Nash demand game, the stag hunt has many interpretations in political and social philosophy. Skyrms (2004) views the transition from the sub-optimal hare equilibrium to the pareto-optimal stag equilibrium as analogous to the transition from the state of nature to an efficient social contract. In a similar vein, Binmore (1994) views the transition from hare to stag hunting as movement from the status-quo to desirable institutional change. For definite social change to occur, however, all must participate in the costly task of bringing said change about.[53]

Unfortunately, the basin of attraction of the hare hunting equilibrium is much larger than the basin of attraction of the stag hunting equilibrium under the replicator dynamics. Additionally, Kandori et al. demonstrate under a stochastic dynamic the population will in the long run spend all of its time at the hare hunting equilibrium. While these pessimistic results suggest collective action is doomed, a number of simple mechanisms encourage widespread stag hunting. Skyrms (2002, 2004), Zollman (2005) and Bruner (forthcoming) have demonstrated the importance of pre-game costless communication in the stag hunt.[54] If individuals are allowed to condition their behavior on the signal of their counterpart, cooperation is highly probable. These

[52] Additionally, there exists a mixed equilibrium where each individual plays hare with a probability 1/3 and stag with probability 2/3. Only the pure equilibria of the game are asymptotically stable.

[53] In this vein, Aaron James (2012) argues that the stag hunt nicely captures the strategic interaction underlying climate change negotiations. Although all countries may prefer abating carbon emissions, no country wants to make such change unilaterally.

[54] Also see Smead and Huttegger (2011) as well as Wagner (unpublished). Also relevant is Wagner (2012) who demonstrates a combined game of the stag hunt and the Nash demand game leads to both more cooperation and fair divisions.

past investigations have worked under the assumption that signals are plastic and are updated just as quickly as one's hunting behavior is. Since we are interested in the dynamics of racism, we first explore the stag hunt played with *fixed* costless pre-game signals, and then later a situation where individuals have both plastic and fixed signals.

We once again consider two populations, the Blues, consisting of blue-skinned individuals and the Reds, consisting of red-skinned individuals. Since individuals cannot control their skin color the ratio of Blues to Reds stays constant. Individuals are randomly paired to play the stag hunt and can utilize one of three strategies: only hunt stag with like-colored individuals, hunt hare unconditionally and hunt stag unconditionally.[55] Table 3.4 lists the results under the two-population replicator dynamics. For a point of comparison, in the one-population replicator dynamics with no signals, approximately 67 percent of simulations arrive at the hare hunting equilibrium while the remaining runs head to the stag hunting equilibrium.

What is immediately evident from Table 3.4 is that racist behavior, where individuals only hunt stag with those of the same race, tends to flourish. States where both populations only cooperate with those of the same race are quite common. Since all employ this race-based strategy no one has incentive to unilaterally deviate. We shall refer to this unfortunate state of affairs as the "racist equilibrium." What also emerge are states where all the members of one population only cooperate with their own race, while all the members of the other population unconditionally hunt hare. In-group cooperation thrives within one race while the second group is stuck at a sub-optimal equilibrium. This state of affairs is also stable and has a considerable basin of attraction.

Just as we did in Section III with the Nash demand game, we adjust the relative size of the two populations. Table 3.4 also presents these results. We see that as one group grows larger in size two significant things occur. First, the number of simulations that head to unconditional hare hunting decreases while the number of simulations that lead to unconditional stag hunting increases. As the minority constitutes a smaller fraction of the overall population attaining the pareto-dominant equilibrium is slightly more likely. Second, the state in which all members of the majority favor in-group members while all in the minority hunt hare

[55] In order to simplify the analysis, we do not consider strategies that only hunt stag with those not of the same color.

unconditionally is more probable as the minority shrinks in size. Individuals in the majority only cooperate with like-skinned individuals while the minority population is stuck at the suboptimal hare equilibrium. Just as we uncovered in the Nash demand game, there are substantial perks associated with being in the majority.

Minority	*Unconditional hare hunter*	*Both pop. are racist*	*Racist maj., Hare min.*	*Racist min., Hare maj.*	*Unconditional stag hunter*
0.5	34.4	15.7	22.6	21.9	5.4
0.4	34.9	15.4	20.3	23.8	5.7
0.3	33.9	14.9	19.2	25.5	6.5
0.2	32.6	15.4	18.2	26.2	7.7
0.1	31.5	14.9	16.1	26.9	10.6
0.05	30.7	14.7	15.0	27.2	12.0

Table 3.5: results of simulations of the two-population discrete-time replicator dynamics. Each entry summarizes the results of 10,000 simulations.

Reducing Racism in Population Games

We now turn our attention to the question of how to reduce racist behavior. As the above results illustrate, it is quite likely that race-based strategies become prevalent in both the stag hunt and Nash demand game. We explore one possible means of reducing the number of such racially charged strategies. This is done by allowing individuals to condition their behavior on the plastic features of their counterpart.

Individuals in this new model send two costless pre-game signals. One of these signals is fixed and, as in the previous section, can take the value of Blue or Red. The new signal we introduce is plastic. Like the fixed signal, a plastic signal is an immutable costless pre-game signal. The main difference, however, is that while fixed signals remain constant, plastic signals are allowed to evolve.[56]

We investigate the simplest case where there are two possible fixed signals and two possible plastic signals, making for a total of four signal combinations. Individuals have the option of conditioning their behavior on their counterpart's fixed signal or plastic signal. In other words, agents can choose to only cooperate with those who share their fixed signal, and

[56] Or if we're thinking of this model as one involving cultural evolution, plastic signals can be easily imitated by others in the population.

likewise, individuals can select to only cooperate with those who possess the same plastic signal.[57] There are 16 strategies in total.[58] We once again investigate the long-term behavior under the two-population replicator dynamics. Results of this simulation are shown in Table 3.5.

Just as in the previous section, there exists a racist equilibrium in which both populations only cooperate with those of the same race. The addition of plastic signals results in a new equilibrium in which *all* individuals adopt the same plastic signal and only cooperate with those who send this signal. Such a configuration guarantees that all individuals hunt stag with each other. Call this equilibrium the "open-minded" equilibrium. Our results indicate that allowing individuals to condition their behavior on the plastic signals of their counterpart does in fact lead to more cooperation and fewer instances in which race-based strategies dominate the population. In fact, the vast majority of simulations end up at the open-minded equilibrium.

Like we have seen before, minorities are more likely to arrive at the sub-optimal hare hunting equilibrium. In fact, the basin of attraction for the equilibrium in which the minority unconditionally hunts hare while the majority hunts stag with in-group members is considerably larger than it was in the case in which individuals just possessed fixed signals. Also note that the basin of attraction of the unconditional hare hunting equilibrium dramatically decrease as the minority decreases in size, a finding that is simultaneously accompanied by an increase in the number of simulations that lead to the open-minded equilibrium. In other words, as the population becomes more homogenous it is easier for the agents to utilize their plastic signals to coordinate cooperative behavior.

We now introduce more plastic signals to our model. We attempt to determine whether the number of plastic signals makes it easier to attain high levels of cooperation. We have good reason to think so. Pacheco et al. (2010) have shown a positive relationship between the number signals and the prevalence of cooperative behavior.

---

[57] Note that we don't allow for strategies that only cooperate with those who send the same fixed signal-plastic signal *combination.*

[58] A strategy is once again a vector with three components. The first entry specifies the fixed signal, second entry specifies the plastic signal and the third entry specifies whether the individual: (i) unconditionally cooperates, (ii) unconditionally defects, (iii) cooperates only with those with the same fixed signal or (iv) cooperates only with those with the same plastic signal. Just as we did before, we exclude "traitor" strategies that cooperate only with out-group members. This is once again done to simplify the analysis.

We expand our above model by allowing individuals to now have three possible plastic signals they could send to their counterpart. When the individuals are given two fixed and three plastic signals the basin of attraction for the open-minded equilibrium increases. The basin of attraction for the open-minded equilibrium continues to grow in size as we continue to add additional plastic signals. Allowing individuals to condition their behavior on malleable features of the individual results in wide-spread cooperation. See Table 3.6.

Minority	Both open-minded	Both racist	Racist min., Hare maj.	Racist maj., Hare min	Unconditional hare hunter
0.5	29.07	9.82	15.33	14.22	31.48
0.4	30.33	9.43	13.07	16.13	30.94
0.3	33.85	9.36	11.46	17.62	27.57
0.2	39.64	9.38	8.86	17.65	24.38
0.1	45.67	8.61	7.13	17.38	20.92

Table 3.6: results of simulations of the two-population discrete-time replicator dynamics. Each entry summarizes the results of 10,000 simulations. Individuals are endowed with both a plastic and fixed signal.

Number of plastic signals	Both open-minded	Both racist	Racist min., Hare maj.	Racist maj., Hare min	Unconditional hare hunter
2	29.8	5.2	14.4	14.1	32.1
3	47.44	4.41	8.54	8.36	31.1
4	60.0	2.9	3.2	4.2	29.7
5	76.6	0.8	2.2	1.8	18.5
6	82.4	.3	.7	.8	15.5
7	87.6	0.0	.5	.4	11.4
8	90.0	0	.2	.1	9.4

Table 3.7: results of simulations of the two-population discrete-time replicator dynamics. Each entry summarizes the results of 10,000 simulations. Individuals are endowed with two fixed signals and we allow the number of plastic signals to change from 2 to 4.

3.6 Discussion

This paper takes some needed steps to better understand the dynamics of discriminatory norms, as well as the impact such norms have on minority groups. We do so by representing both race and racism in a game-theoretic setting through the use of costless pre-game signals. With this formalization we then turn our attention to a variety of strategic games to better

understand how race impacts bargaining behavior as well as collective action. Our findings are two-fold. First we uncover that minority status often leads to a weak bargaining position. Members of the minority are frequently the victim of discriminatory norms that leave them with a fraction of what members of the majority receive. This however, needn't always be the case, and we briefly explore a game in which the majority is less likely to get their favored result. Second we discover a promising means of reducing racist behavior in population games. Namely, we allow individuals to condition their behavior on plastic features of their counterpart. We attend to each of these findings in turn.

First recall that in both the Nash demand game and the stag hunt game the minority was more likely than the majority to be the victim of a discriminatory norm. Note, however, that this "minority disadvantage" was simply in virtue of the group's size. It was not the case that the racial groups differed in some systematic way – both valued the disputed resource equally, and both derived the same utility from coordinating on the pareto-dominant equilibrium in the stag hunt game. Differences in group size seem to be enough to consistently bring about glaring economic inequalities. This finding is of interest when one considers that social psychologists have documented that many believe economic disparities between races are in large part due to differences in effort, values and work ethic.[59] Many American individuals, for example, were inclined to say that economic differences between whites and blacks simply reflected the unwillingness of African-Americans to work to better themselves and take advantage of opportunities bestowed upon them. This paper uncovers a plausible alternative explanation that can account for such glaring inequalities. Differences in group size translate into different levels of inertia, which in turn result in differential economic outcomes.

This paper also highlights the role of inertia in the evolution of norms. Minorities were at a distinct disadvantage in both the Nash demand game and the stag hunt because they rapidly adapted to the initial behavior of the majority. Work by both economists (Young, 1992, Gallos, forthcoming), and biologists (Bergstrom and Lachman, Nowak as well, right?) have linked inertia to equilibrium selection as well, suggesting the role of inertia is not just an artifact of our model but instead points to a more general story about the nature of inertia in strategic settings.

[59] See for example, Henry and Sears (2002) and the related literature on so-called "symbolic racism."

We also explore a means of reducing racist behavior. The question of how to reduce racist behavior is an important one, and has been discussed in one form or another since at least Friedman. Yet little progress has been made over the past half century. Schelling's famous neighborhood model, as well as similar models, are incapable of uncovering the key to racism reduction in part because Schelling artificially holds fixed the racial preferences of his agents. That is not to say there is no work on the topic of racism reduction. Grim et al. (2005) explores the "contact hypothesis" with use of computer simulations and uncovers that if individuals are embedded on a social network, desegregated networks promote race-blind strategies and wide-spread cooperation. We explore an alternative route, and instead try to induce individuals to focus on other salient (and plastic) attributes of their counterpart. This is a novel proposal which to the best of our knowledge, has not been discussed by psychologists, policy makers or social philosophers.[60] Allowing individuals to condition their behavior on plastic traits tends to move people away from racist behavior because plastic signals can coordinate cooperative behavior in the stag hunt game with lightning speed. Thus Grim and I provide two distinct means of expunging racist behavior from the population. Yet while desegregation can promote color-blind behavior, the process of desegregating may unfortunately come at a rather steep cost. Converting a segregated neighborhood into a non-segregated neighborhood is both financially costly and politically risky. My approach, on the other hand, side-steps these high costs. If anything, all my approach needs to be successful is for the plastic traits of individuals to become more salient to the agents. Once individuals learn to condition their behavior on plastic traits, cooperation is nearly unavoidable.

Finally, one may think it peculiar that our suggestion to avoid discrimination based on race is to instead encourage people to focus on and discriminate based on *other* features of the agent. As we saw, the so-called "open-minded" equilibrium involves all adopting the same plastic signal and all refusing to cooperate with those who do not send this plastic signal. If these signals are, for example, the color of the shirt the agent wears, then this homogeneity is a small price to pay. If, however, the malleable signal is to display a particular religious emblem on one's clothing then the open-minded equilibrium becomes much less appealing. Plastic

[60] See Kelly, Faucher and Machery (2010) for a review and assessment of the social psychology literature on racism as well as three popular proposals to reduce racist behavior.

signals facilitate cooperation between those with different fixed signals, yet this requires that all adopt the same plastic signal. Is it possible to eventual re-introduce signal diversity into the population or are we stuck with this homogeneity? Recovering signal diversity is possible but requires a slightly more complicated (and slightly different) model. We briefly outline such a model, based on work by Bruner (manuscript), below.

Consider a visible signal, the value of which can be thought of as a number on the unit interval. When deciding whether to cooperate or not in the stag hunt game individuals first determine how much distance lies between their two signals. If this value is less than the agent's "tolerance level," the individual behaves cooperatively in the stag hunt game. Individuals imitate the tolerance level and signal of more successful peers, and via computer simulation we can track the long-term behavior of the population. Now consider a slightly more complicated version of this set-up where individuals possess two visible signals, each represented as a number between one and zero. Individuals can now be thought of as inhabiting a point in a two dimensional space. Additionally, one of these signals is plastic while the other is fixed. Individuals still have a tolerance level, but are further classified as open-minded or racist. Racist individuals only care about the fixed trait of their counterpart. When determining whether to cooperate with a counterpart they measure how much distance lies between their fixed traits, and if this number is less than their tolerance level, they cooperate. Open-minded individuals engage in a similar procedure except with respect to the plastic trait. Individuals still imitate the tolerance level and plastic signal value of those more successful than themselves, but additionally match whether this individual was racist or open-minded.

What we get is something quite striking. Wide-spread cooperation is possible due to the plastic traits. Plastic traits help facilitate cooperation among a handful of agents, which in turn means the plastic trait and tolerance level of these successful agents are slowly adopted by more and more of the population. Soon all are open-minded and, as we saw in section V, all adopt the same plastic signal. Thus allowing individuals to condition their behavior on either the fixed or plastic signals of their counterpart generates high levels of cooperation and homogeneity (with respect to plastic traits), just as we observed in section V. This, however, is not the end of the story. If we introduce mutations – i.e., if we assume that with a small chance either the value of an agent's plastic signal or their tolerance level is slightly perturbed – signal diversity can easily

be reclaimed. This dynamic is illustrated in Figure 3.1 and a similar sort of dynamic is uncovered in the previous chapter.[61] The logic behind this is simple: there is a weak selection for increasingly tolerant agents. Mutants who become too intolerant do exceptionally poorly – they forego stag hunting opportunities with the majority of the population. The increase in tolerance then allows for slightly more signal diversity, which in turn selects for even slightly more tolerant agents. The steady increase of tolerance allows us to recover the diversity that was initially present in the population.

[61] We ran a number of simulations of this model. The details are as follows: 200 agents were randomly assigned both a fixed and plastic signal (from zero to one) and a tolerance level (from zero to 0.25). Individuals were paired to play the stag hunt game ten times before being paired to imitate. Individuals imitated the tolerance level and plastic signal of their counterpart as well as their racist or open-minded disposition. We varied the initial number of racists in the population. Of the 100 simulations ran in which half of the initial individuals were racist all 100 of these resulted in the dynamic illustrated in Figure 3.1. When 60 percent of the initial population was racist 94 of 100 simulations resulted in the dynamic illustrated in Figure 3.1. The remaining simulations resulted in all 200 individuals adopting the racist disposition. When 70 percent of the initial population had a racist demeanor 51 of 100 runs resembled Figure 3.1 and the rest resulted in all adopting racist dispositions.

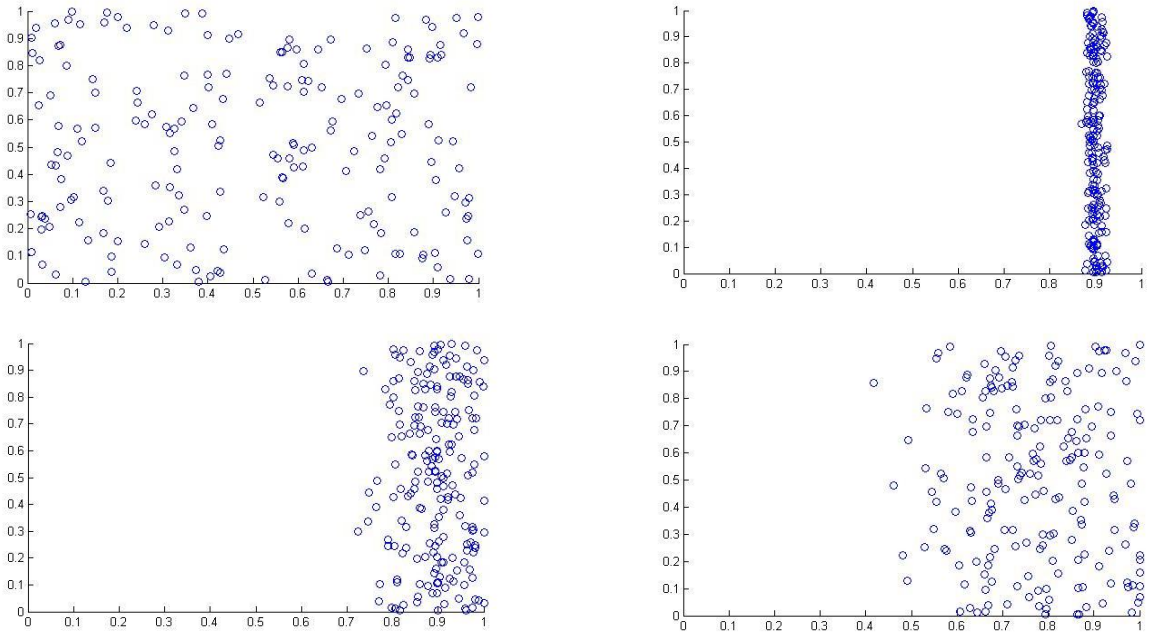

Figure 3.1: Simulation results for the two-trait case.

Chapter 4

THE POSSIBILITY OF PLURALISM

4.1 Introduction

Consider what is often referred to as political pluralism, the doctrine that begins with the observation that there exist many conceptions of the good, and contends individuals should be in some sense tolerant of their peer's moral beliefs and practices. Political pluralism has generated much heated discussion regarding, among other things, the role of religious values and reasoning in public discourse and how to differentiate so-called reasonable comprehensive doctrines from the lot. Yet equally important is the somewhat more practical concern regarding stability. What sorts of arrangements are sustainable in the face of what Rawls referred to as the "fact of pluralism"? In the course of this paper, we attempt to determine in what sense, if any at all, a pluralistic and tolerant society can resist episodes of intolerance.

We have, of course, good reason to be doubtful that tolerance can be sustained for long periods of time. Tolerance can easily wane, as evidenced by Europe's response to the recent influx of Muslim immigrants. Negative attitudes toward Muslims have resulted in discrimination in both the workplace as well as in hiring practices. Polls conducted by Amnesty International in France and Germany find that the majority of respondents believe Islam is simply not compatible with their country's way of life. It is not hard to see how these attitudes can be easily transformed into formal policies, as in the case of Switzerland, where a constitutional amendment has outlawed the construction of minarets.

Recently, game-theoretically minded political philosophers have begun to investigate how tolerant behavior can be sustained in a diverse society. Muldoon et al., for example, show high levels of tolerance are attainable when there are gains to be had from interacting with those different from oneself. Gaus (2011) has argued that a system of social moral rules chosen by the "hand of cultural evolution" allows for peaceful cooperation in a pluralistic society and can be sustained by agents who take it upon themselves to punish rule-violators. Gregory Kavka, in a little known paper "Why Even Morally Perfect People Would Need Government," tackles our problem head-on, and examines a scenario where disagreement is inevitable since agents subscribe

to different conceptions of the good and must settle such disputes without the help of a third-party. Kavka's conclusion is that tolerant attitudes will inevitably erode, demonstrating that there is a sense in which a tolerant pluralistic society is unstable. Without some sort of central authority ensuring that conflict does not arise, agents who stick by their particular conception of the good and refuse to compromise with others, can, Kavka argues with the help of some evolutionary models, infiltrate a population of compromising and tolerant agents. We explore this line of thought further and argue that Kavka's analysis is slightly off base because the most natural way to interpret the strategic situation faced by our agents is not as a prisoner's dilemma, but as a conflictual coordination game. Running with this insight, we develop our own model of this situation and determine that under many conditions, pluralist agents who are tolerant of their peer's different moral beliefs and practices can thrive.

Lastly, we consider situations of homogeneity – communities in which all share the same conception of the good. Disagreement can nonetheless still arise due to the fact that agents must, without the help of a governmental agency or third-party, behave as "judges in their own case." Self-bias can thus be a source of conflict even when agents share values. Modeling this once again as a conflictual coordination game, we show that disagreements can be peacefully resolved and conflict avoided under realistic circumstances. We spell out what this means for Robert Nozick, whose famous account of the evolution of the minimal state takes as its starting point such a setting. We argue that the results of this paper as well as other findings in the evolutionary game theory literature suggest that the transition from the state of nature to the minimal state may not be as inevitable as Nozick had suspected.

4.2 Pluralism, Conflict and Compromise

We begin with a community of agents who do not necessarily have the same conception of the good (throughout we use "conception of the good," "comprehensive doctrine," and "value system" interchangeably). This in turn means our agents may disagree on substantive moral issues. How individuals behave when such disagreements arise in their everyday dealings with others is our primary concern. Kavka stipulates that tolerant, so-called pluralist, agents would place a high premium on arriving at a peaceful compromise when confronted by disagreement.

These agents will desire to settle disputes before they become too disruptive as to threaten the tranquility of the community.[62]

Thus there is little to no conflict in a community of pluralists because whenever disagreement does arise those involved will settle by coming to some sort of compromise. Kavka envisions the two agents literally "split the difference" between their value systems in order to arrive at a mutually acceptable resolution. One may wonder how this is done, or even if this can be done in many circumstances. While I concede that splitting the difference may not make sense in certain settings, we can, for example, have agents instead flip a coin to determine which of them get their way.[63] On average, then, this strategy may be similar to the sort of compromise Kavka had in mind.[64] Either way, disagreement spurred by diversity is kept in check and not allowed to transform into outright conflict. Yet is such an arrangement stable? Kavka provides us with good reason to think that our pluralist society is quite fragile.

Kavka's explanation for why a population of pluralists is not stable rests on some already established results from evolutionary game theory. Evolutionary game theory concerns itself with the study of behavior in strategic settings. In particular, individuals are randomly matched to play a game with other members of their community, and the composition of the population changes over time as determined by the relative success of the various strategies utilized in the population. Thus in the prisoner's dilemma, for example, if agents who defect do better on average than those who cooperate, then over time increasingly more people will begin to utilize defect. Standard models of evolutionary change such as the replicator dynamics are often interpreted as modeling

[62] Thus part of the reason to value compromise so much is a fear of reverting back to the Hobbesian state of anticipation. Yet this is of course not the only reason to value compromise. Eric Cave (1996), for example, provides a fuller and more Rawlsian version of Kavka's compromiser as an agent who adopts the agreed upon public conception of justice, and desires to exercise this sense of justice as well as help foster a community in which all members are moved to act in a way that is consistent with this publically-shared sense of justice

[63] Either way, Kavka assumes that while pluralist agents may have first-order disagreements, they do not have second-order disagreements – i.e., they do not quarrel over how, exactly, to "split the difference."

[64] Following Cave (1996), another way to conceive of this is that our pluralist agents settle their dispute by disregarding their personal comprehensive doctrine and instead appeal to the public conception of justice already established in the population. Of course the question of how this public conception of justice came into existence is a discussion for another paper. See, for example Gaus (2011) who suggests some sort of cultural evolution could lead to one of a whole set of reasonable justifiable comprehensive conceptions coming to the front and acting as the social morality that guides action in the public sphere.

cultural evolution, where individuals imitate the behavior of those more successful than themselves.[65] For the remainder of this paper, we will use the term evolution to refer to this sort of imitation-based cultural evolution.

Now return to our population of pluralists. Consider an agent who when confronted by disagreement does not make any attempt to find common ground and instead stubbornly demands his counterpart respect the dictates of his comprehensive doctrine. Call such an agent an *uncompromiser*. When an uncompromiser meets a pluralist, the pluralist, valuing compromise above all else, is willing to concede and let the compromiser have his way.[66] Of course the downside to being an uncompromiser is that when two uncompromisers meet the result is at best a stalemate, with neither willing to budge, and at worse can escalate to a physical confrontation. Kavka models this interaction as a prisoner's dilemma, where uncomrpomisers are thought of as defectors and pluralists as cooperators. One now has reason to be an uncompromiser because uncompromisers do better than pluralists against both pluralists and uncompromisers. It is therefore easy to see why our liberal pluralistic society is not stable – uncompromisers can easily invade and will slowly drive the pluralists extinct.

This is of course not to say that there is no hope for our pluralists. Consider, as Kavka does, more sophisticated strategies that utilize information about the past play of their counterpart. Such "reputation-based" strategies have the potential to keep uncompromisers at bay. For example, Kavka suggests an "Avenger" type who is inclined to initially compromise and continues to do so as long as their partner compromises with them (this is essentially the so-called grim trigger strategy). Another more complicated strategy, "Guardian," only compromises with those who in all of their previous interactions have themselves compromised. Interestingly, similar sorts of reputation-based strategies have been thoroughly studied in economics, and Guardians have

[65] See Sandholm (2010) for more on this, as well as a demonstration that the "imitate the more successful" dynamics is formally equivalent to the replicator dynamics.

[66] Of course we could imagine a less extreme sort of compromiser who compromises with those possessing a similar disposition and digs her heels in when confronted with uncompromisers. Kavka does not entertain such a strategy and all we will do in this paper is register that such a strategy is very well possible. Instead we will focus on how a society of extreme compromisers could ever be viable.

been shown to be a viable means of ensuring cooperation in a variety of settings.[67] Yet this too is problematic because (i) these models are not evolutionary in nature and thus do not involve the low-rationality agents Kavka has in mind and (ii) these models are known to be very fragile, and it has been shown that cooperation in these reputation based models unravel if certain assumptions about the distribution of information are relaxed (see, for example, Vanderschraaf, 2007).[68]

Instead of attempting to tweak these models of reputation-tracking we instead suggest that the prisoner's dilemma is not the most natural way to capture the strategic situation faced by two agents in the process of working through a disagreement. In particular, it may often be in the interest of both agents to somehow resolve the disagreement, thereby avoiding a stalemate. This line of reasoning is particularly appealing if we assume that our agents are engaged in some sort of collaborative or cooperative interaction that both stand to gain from, regardless if the 2disagreement is resolved on their terms.

To make this more concrete, consider an example provided by Gerald Gaus. In Gaus (2011) the following "Kantian coordination game" is presented. In this case individuals must decide between a number of social moral rules that could potentially govern day-to-day interactions.[69] As is the case in our pluralist community, there is disagreement as to which system of social moral rules is best. Which system of social moral rules is ultimately adopted by the population is, according to Gaus, determined by some sort of cultural-evolutionary process. In the simplest case, two agents must determine whether to use social rule A or social rule B when interacting with each other. The game is a coordination game because presumably individuals all benefit if their dealings are governed by the same system of moral rules – i.e., the social rules, when utilized by both agents, are able to successfully facilitate cooperation and help the two agents

[67] Kandori (1992) is a classic paper in this literature. Skyrms (1995) nicely summarizes many findings while demonstrating their significance to political philosophy.

[68] See Cave (1996) for more on reputation-based strategies in the context of a pluralistic society. Vanderschraaf (2007) suggests that although decentralized communities of agents face informational problems a minimal sort of government could be tasked to keep track of reputations. One may also appeal to the more recent literature on costly or altruistic punishment. Boyd, Gintis, Bowles and Richerson (2003) have shown punishment administered by individuals in a decentralized community can be used to stabilize social norms in a variety of settings.

[69] Social moral rules are very similar to what we have been referring to as comprehensive doctrines, with the nuance that these social moral rules can somehow be justified to all others in the population. Gaus believes that this puts some structure on the content of these rules, but not enough to narrow them down to just one social morality.

coordinate their behavior. Conflict and misunderstandings are minimized. It is nonetheless a conflictual coordination game because not all have the same preferences over possible social moral rules. Consider, for example, a situation involving two agents, Annie and Betty, who prefer social rules A and B, respectively.

	A	**B**
A	3, 2	0, 0
B	0, 0	2, 3

Table 4.1: Conflictual coordination game between Annie (row player) and Betty (column player). This game, taken from Gaus (2011) is known in the game-theoretic literature as the Battle of the Sexes and the two pure-strategy equilibria are (A, A) and (B, B).

While Annie of course prefers that their behavior is both governed by A, she considers the situation where they fail to coordinate on a social morality as inferior to the situation where they both utilize B. Likewise for Betty. This situation is represented as a payoff table in Table 4.1. Failing to coordinate on a social morality is particularly devastating. Now consider the situation in which Annie is an uncompromiser – she refuses to play anything but her preferred social rule, A. If Betty is equally stubborn the two of them do particularly badly. Yet if Betty instead decides to be accommodating and is willing to play A when paired with the stubborn Annie, they successfully coordinate. While Betty ends up not doing as well as Annie, she nonetheless does better than how she would have if she had refused to concede. Finally, assume that if both Annie and Betty desire a compromise, they flip a coin to determine whether they will both play A or B (which yields a payoff of 2.5).

This interaction is nicely summarized in Table 4.2, and it should be noted that this game is essentially the more familiar hawk-dove game. In the hawk dove game, a precious resource is contested, and agents can either aggressively pursue the resource at all costs, or back down and concede the resource when conflict seems imminent. In this two-player, two-strategy simultaneously move game there exist two pure Nash equilibria. Both equilibria consist of one agent aggressively demanding her way (hawk) while the other acquiesces (dove). There also exists a mixed Nash equilibrium in which both agents play dove and hawk with a certain probability. It is easy to see why the pure-strategy Nash equilibria are stable: since the cost of conflict is high, one does best to acquiesce against a partner who is aggressive. Likewise, being aggressive is the best

course of action if one's counterpart is dovish. Thus just as in the prisoner's dilemma, a situation only involving pluralists is not stable – each has incentive to unilaterally deviate. Yet unlike the prisoner's dilemma, a population of uncompromising agents is likewise unstable. Pluralists don't get their way, but avoid the cost of conflict that the majority of uncompromisers incur.

Thus pluralists seem to be at less of a disadvantage when we move from the prisoner's dilemma to the hawk-dove game. Yet as indicated above a community of pluralists is still fragile, and can be easily invaded by uncompromisers. The hawk-dove game is an apt representation of the scenario we are concerned with if both agents stand to benefit from successfully avoiding conflict. While this of course does not cover all possible cases, it does speak to those very important circumstances where agents are engaged in a joint-project whose completion benefits all involved. Our aim in the next section is to better understand and identify the conditions under which high levels of tolerant, dovish behavior can be sustained. Before we begin, however, one more note must be made about the underlying game. Observe that when two agents share the same value system there is little to no disagreement on substantive moral issues and thus no possibility for conflict.[70] In other words, when interacting with another who shares your values no strategic considerations are necessary. Yet in the case where two agents subscribe to different conceptions of the good, we revert to the hawk-dove game. Thus, an agent does not necessarily play the same game with all else in the population.[71] We define the Kavka-conflict game similarly. When two agents have the same comprehensive good they simply receive the payoff of V, and when they subscribe to different comprehensive goods they play the hawk-dove game seen in Table 4.3.[72] This new game will be further explored in the next section.

[70] This assumption will be relaxed in section IV.

[71] The "game" played when two agents possess the same comprehensive doctrine is not really a game at all, since both receive the same payoff regardless of whether they are a compromiser or not.

[72] Kavka himself failed to note that conflict is avoided if agents subscribe to similar value systems. Note that while the Kavka-conflict game is a game of complete information, it can be also conceived of as a game of incomplete information, and that the two variants of this game are formally equivalent.

	Pluralist	**Uncompromiser**
Pluralist	2.5, 2.5	2, 3
Uncompromiser	3, 2	0, 0

Table 4.2: Hawk-dove game.

	Pluralist	**Uncompromiser**
Pluralist	V/2, V/2	0, V
Uncompromiser	V, 0	(V-C)/2, (V-C)/2

Table 4.3: Generic payoff table for the hawk-dove game (C > V).

4.3 Model and Results

In this section we explore both the hawk-dove game as well as the Kavka-conflict game. We see that there are a number of ways to minimize conflict in the population, but these means are often undesirable because they either (i) result in homogeneity (i.e., only one comprehensive doctrine is observed in the long-run) or (ii) result in inequalities that systematically disadvantage the adherents of one comprehensive doctrine. We avoid both of these drawbacks if we instead embed agents on a social network and allow for somewhat sophisticated strategies such as tit-for-tat. This ensures that pluralists thrive, and we see that tolerant behavior is actually made all the more likely as the population becomes increasingly diverse.

Let us first begin with the standard hawk-dove game. In this case agents can either behave aggressively or acquiesce. Under the replicator dynamics, a population of agents will head to a polymorphic equilibrium that consists of both hawks and doves, and all sorts of interactions are possible in this mixed state.[73] Sometimes pluralists interact with other pluralists, in which case a compromise is struck, while at other times pluralists encounter uncompromisers and acquiesce. Pluralists are common but by no means the norm. Additionally, conflict occurs quite regularly, as uncompromisers often run into fellow uncompromisers.

As many have long noted, conflict in the hawk-dove game is avoidable if individuals in the hawk-dove game can condition their behavior on a visible feature of their counterpart.[74] Skyrms

[73] This equilibrium is asymptotically stable. The exact proportion of hawks and doves found in equilibrium corresponds to the mixed Nash equilibrium.

[74] More generally, if the agents can condition on some sort of a public signal then they can coordinate on one of the available pure-strategy equilibria in the hawk-dove game. This is an instance of a correlated equilibria, first pioneered by Robert Aumann.

and Zollman (2011) provide a nice example of this for the hawk dove game.[75] Consider two populations of agents, the Greens and the Blues. Assume that before play each agent can correctly determine from which population their counterpart hails from, and can condition their behavior on this information. The following arrangement is now stable: whenever a Green agent meets a Blue agent the Green agent plays dove while the Blue agent plays hawk. This means that interactions between Green and Blue actors are conflict-free – Blues and Greens never simultaneously play hawk. Of course the kicker is that the poor Green population continually receives a smaller payoff than those in the Blue population.

Shifting our attention to the Kavka-Conflict game, it is evident that asymmetries can play a similar role. Instead of Blue and Green types, an agent subscribes to one of two comprehensive doctrines. Agents can then condition their behavior on the comprehensive doctrine of their counterpart and imitate the hawkish or dovish behavior of their more successful peers. As we saw above, this leads to little conflict, but results in the members of one value system always conceding when they interact with out-group members. Thus, contra Kavka, outright conflict (cases where two uncompromisers butt heads) is avoided, but things become rather one-sided.

Note that in the above model we've assumed that agents only update whether they are a compromiser or uncompromiser, and not their comprehensive doctrine. While I believe this modeling decision is made with good reason – agents are much more likely to update how they deal with disagreement than make substantive changes to their comprehensive doctrine – it is an assumption that can nonetheless be relaxed, and we see that when both doctrine and compromising behavior can be imitated, conflict is once again easily avoided. Unfortunately, this comes at the cost of diversity – the end result is a population in which all adhere to the same comprehensive doctrine and nearly all are unwilling to compromise when confronted by disagreement.[76]

Thus while we have shown something quite interesting – that conflict can in many cases be completely avoided – we have failed to demonstrate that this is due to the population as a whole

[75] Sugden (1986) also has a nice example of this with an application to state of nature theory. The first instance of this I know of in an evolutionary context is John Maynard Smith's work.

[76] It is easy to see why the population lacks pluralists. Consider the situation in which adherents of doctrine A refuse to compromise and adherents of doctrine B always compromise. If agents are now allowed to update their doctrine followers of B will slowly imitate both the doctrine and behavioral disposition of the adherents of A.

adopting tolerant attitudes. To further investigate the Kavka-conflict game we utilize an agent based model in which individuals are embedded on a social network. It has long been realized that spatialized games often generate more cooperative behavior than those involving a well-mixed population.[77] Individuals are now placed on a 64 by 64 lattice which wraps around the edges (effectively making this a torus). Individuals play the Kavka-conflict game with all the individuals in their von Neumann neighborhood, using the same strategy against each agent. Individuals then survey their neighbors to determine whether any neighbor received a higher payoff than themselves. If so, the agent adopts the strategy of this most successful neighbor (in the case of a tie between neighbors, which neighbor is imitated is determined by chance). If not, the individual continues with her present strategy.[78]

We can of course modify this update protocol in a number of different ways. For example, agents can imitate probabilistically proportional to fitness. Likewise, we can also vary what is typically imitated. We'll first examine the case in which agents imitate both the value system and strategic behavior of their neighbor. We'll later examine the realistic case in which an agent holds convictions as fixed and only change whether they compromise or not.

Figure 4.1 outlines what a typical simulation run looks like. Recall that in this baseline case individuals imitate both the compromising behavior as well as the comprehensive doctrine of their more successful neighbors. The results reflect this fact: the population is carved up and organized *by value system*. That is to say, there are large regions where only one comprehensive doctrine prevails. For the most part individuals are uncompromisers and this is especially true on the border *between* different doctrines.[79] Since the population is segregated the amount of overall

[77] See Pollock (1989), Nowak and May (1992), Grim (1995). However it should be noted that in our game of interest, the hawk-dove game, the opposite result holds. Embedding agents on a network often results in less cooperation than in a well-mixed population (Hauert and Doebeli, 2004).

[78] This update-rule, imitate-the-best, is commonly used in evolutionary games (see Skyrms, 2004).

[79] This result is reminiscent of Zollman (2005), who investigates a similar model. Agents are once again embedded on a lattice, assigned to one of two "groups" and tasked to play the stag hunt with their neighbors. Individuals have the option of conditioning their behavior on the group membership of their counterpart. Zollman discovers that although wide-spread cooperation is almost guaranteed, the population is highly segregated.

conflict in the system as a whole is very low. Conflict is relegated to the borders, and there are a sprinkling of compromising agents in the interior.[80]

Once again we see that by and large conflict can be avoided in a diverse population. Unfortunately, this is not due to the fact that agents are more willing to peacefully resolve their disagreements, but instead due to agents segregating themselves in some fashion.[81] Agents avoid conflict by converting to the moral theory that is most prevalent in their neighborhood. The end result is a conflict-free co-existence, but this is achieved by avoiding interactions which may result in disagreement, and thus conflict.[82]

What occurs if we treat comprehensive doctrines as fixed? What we see in this case is that conflict abounds, as over 68 percent of the final population is an uncompromiser. Since enclaves are not possible, the population is characterized by intense conflict as those subscribing to different comprehensive doctrines are forced to interact repeatedly.[83] We find that increasing the number of distinct comprehensive doctrines in the population has the effect of slightly reducing the number of hawks present in the network. When there are five comprehensive doctrines, for example, the proportion of hawks decreases now to 57 percent. Although there are still plenty of hawks in the population, the inclusion of additional value systems seems to have the effect of slightly decreasing the number of aggressive types in the population. This effect is even more pronounced when we introduce more sophisticated strategies, such as tit for tat.

We now allow for tit-for-tat. Individuals now interact with their neighbors repeatedly in-between imitation periods. Those playing tit-for-tat start out by initially compromising, and reciprocate this behavior into the future if their counterpart likewise compromises. However, if they encounter an individual who refuses to budge, they refuse to compromise in future rounds.[84]

---

[80] These borders are stable because while agents on the border do rather poorly, they nonetheless are situated next to agents with the same comprehensive doctrine and strategy that do very well since they are surrounded by similar minded agents.

[81] Approximately 95 percent of the final population was an uncompromiser.

[82] Note that this society is not particularly tolerant in the sense that there is a high chance that any two randomly selected individuals will fall into disagreement and not be able to come to a peaceful compromise.

[83] Note, however, that the proportion of uncompromisers or hawkish players in the Kavka-conflict matches the proportion of hawkish behavior in the traditional hawk-dove game on a lattice.

[84] Of course tit-for-tat is most known for being the most successful strategy in Axelrod's famous computer simulation contest (Axelrod, 1984). Tit-for-tat, however, has been shown to not

In the prisoner's dilemma, tit-for-tat has been shown to thrive in spatialized settings. Grim (1995) demonstrates that even more cooperative variants of tit- for-tat, the so-called generous tit-for-tat,

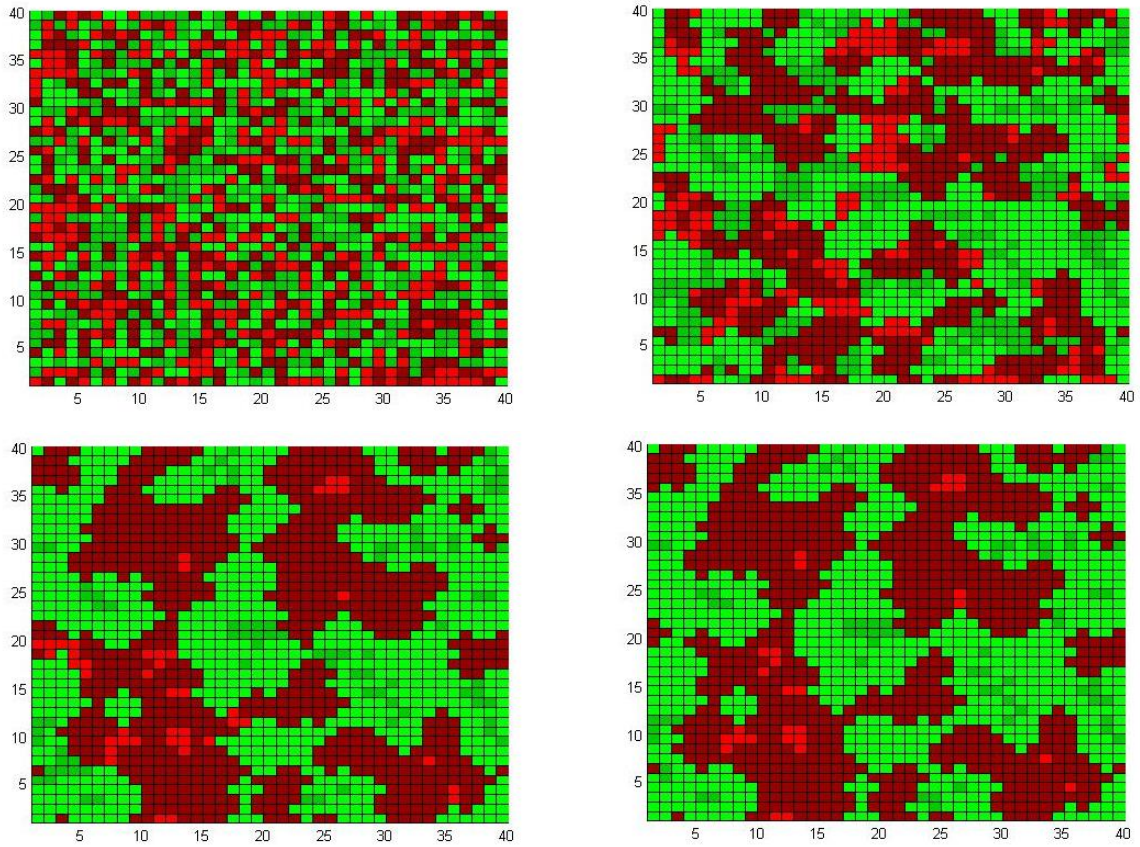

Figure 4.1: Kavka-conflict game on a lattice with two distinct value systems (green and red). Bright red and dark green represent compromisers.

can be sustained when on a network as well. To the best of this author's knowledge, there exist no analogous studies of tit-for-tat and spatial structure for the hawk-dove game. Continuing with the above lattice network, we first allow agents to once again imitate both the comprehensive doctrine and strategic behavior of their more successful peers. We uncover more of the same – we once again encounter the familiar fractionalization we see in Figure 4.1 and the majority of agents are hawks with only a minority utilizing dove or tit-for-tat. However, if individuals are only able to

---

produce such great results in a mixed population of agents when competing in the prisoner's dilemma against always cooperate and always defect (Young and Foster, 1991). That said, when we move to a network tit-for-tat does substantially better, and it can be shown that at least in the prisoner's dilemma, cooperation is possible via tit-for-tat.

update their strategic behavior and not their comprehensive doctrine, we find that surprisingly cooperative behavior thrives. In a community where there are two distinct comprehensive doctrines, 30 percent of the population plays hawk, while 37 percent play tit-for-tat and the remainder utilize dove. The proportion of agents playing tit-for-tat only increases as the number of distinct value systems in the community increases. In the case where there are five distinct comprehensive doctrines in the population, the proportion of tit-for-tat players skyrockets to 55 percent while the number of agents employing dove goes to 32.5 percent.

The above results supports a rather counter-intuitive thesis, namely, that tolerant behavior is spurred in situations where individuals with permanent differences are in some way forced to repeatedly interact with one another. We find, somewhat surprisingly, that in such a case individuals can avoid conflict with each other, and the more diverse the population, the more likely a peaceful coexistence is possible. Note that as the number of distinct value systems in the population grows, the Kavka-conflict game on a lattice with a finite neighborhood becomes increasingly indistinguishable from the traditional hawk-dove game on a lattice. Consider, for example, the limiting case where each agent has her own unique value system. This is essentially the hawk-dove game on a lattice, for no pair of neighbors subscribe to the same comprehensive doctrine. We find that in such a situation hawk is quickly driven to extinction and a combination of tit-for-tat and dove remain, with the vast majority of agents utilizing tit-for-tat.

4.4 Hawks and Doves in the State of Nature

So far we have made one rather large idealizing assumption – that conflict cannot occur between two agents who subscribe to the same comprehensive doctrine. Yet upon reflection it is clear that this need not be the case at all. Well-intentioned agents may not have access to the same facts, or if they do, could have different interpretations of said facts. There may also be disagreement as to how to interpret and apply a moral rule in a particular instance. Additionally, individuals may allow self-interest to clout their judgment. Locke emphasizes this last point, stating, in reference to the law of nature, that even in situations where the law of nature is both "plain and intelligible to all rational Creatures" men are nonetheless "biased by their interests." Such self-bias leads to disagreement that can in turn easily escalate to violence.

Can any of the results from the previous section be applied to the special case in which all share the same comprehensive doctrine? This question can be repackaged as whether or not conflict will arise in conditions approximating the Lockean state of nature.[85] As it is often interpreted, the majority of agents in Locke's state of nature abide by the law of nature, thereby respecting their peer's life, liberty and property.[86] Secondly, agents do not have a "common superior on earth to judge between them" when problems do arise. Agents must take it upon themselves to determine when violations have occurred, and utilize their "executive power" to determine and administer a proper punishment to those who violate the law of nature. Yet if agents are biased in their own favor, as Locke himself assumes, then both parties are prone to exaggerate their claims. The victim will overstate the offense (and thereby overstate how much compensation she deserves), while the guilty party will downplay the damages her actions have caused. Self-bias results in disagreement, and it is unclear whether conflict-free resolutions are possible.[87] Locke is not optimistic, and contends that conflict stemming from such disagreement will cause much "confusion and disorder."[88]

Nozick (1974) takes as his starting point the Lockean state of nature and similarly assumes that agents are self-biased and that conflict spurred by the resulting disagreement is both inevitable and unavoidable, making the "private and personal enforcement of one's rights lead to feuds, and to an endless series of acts of retaliation and exactions of compensation." This in turn marks the beginning of his infamous invisible hand story regarding the emergence of the minimal state.

---

[85] Note that most of the extant literature on conflict in Locke's state of nature focuses on the more Hobbesian question of anticipation (see Kavka, 1986). In this situation one is concerned with exploiting and being exploited by others, and it has been argued by some (Vanderschraaf, 2006; Simmons 1989, 1993) that in these circumstances Locke thinks a state of peace is possible. I am less concerned with exploitation and more concerned with how disagreements due to self-bias are resolved.

[86] See, for example, Nozick, 1974 pg 336 n.10.

[87] In fact Locke seems to believe that conflict is almost inevitable and government is necessary if one desire to avoid the violence and disorder caused by disagreement in the state of nature: "it will be objected that it is unreasonable for men to be judges in their own cases, that selflove will make men partial to themselves and their friends. And ill nature, passion and revenge will carry them too far in punishing others. And hence nothing but confusion and disorder will follow, and that therefore God hath certainly appointed government to restrain the partiality and violence of men" (Locke, Second Treatise 2.13).

[88] Therefore, he urges us to not rely on individual "private judgment," but instead allow the "community [to become] Umpire."

Since judging in one's own case appears to lead to high levels of conflict, there is a demand for private protection agencies tasked to protect rights as well as seek compensation for past violations. Due to economies of scale, as well as competition between firms, Nozick contends that one firm will in the long-run emerge as the so-called dominant protection agency, and this dominant firm will be the bedrock of the minimal state. In other words, through a sequence of interactions involving agents simply attempting to curtail rights violations and ensure their safety, agents in Locke's state of nature will naturally "back into the state."

I believe there is a slight misstep here. Both Nozick and Locke assume that when disagreements arise there is essentially no possibility for compromise – i.e., (potentially violent) conflict is essentially inevitable. Yet as we've seen from the previous section, if we envision disagreement as a conflictual coordination game, conflict is not the only option, and a peaceful settlement is possible. If interactions in the Lockean state of nature are most naturally modeled as a conflictual coordination game such as the hawk-dove game, then we may have very good reason to be skeptical of the claim that conflict is essentially unavoidable in these circumstances.

It is easy to see how the hawk-dove game can nicely capture the strategic situation encountered by agents in Locke's state of nature. Since both parties suffer from bias, they will both overstate their case. If one gets her way then this lucky individual does exceptionally well, while her counterpart is treated rather severely. However, if neither is willing to concede, then, as Nozick claims, a conflict ensues which in turn lead to a fierce feud. Now consider two possible strategies agents can employ. On the one hand, an individual can be steadfast and stand by her initial demands. Alternatively, individuals can favor compromise, and be willing to accommodate their counterpart so as to avoid a potentially violent confrontation. Two compromisers will find some amicable middle ground, and although both do not get exactly what they desire, conflict is avoided. On the other hand, when a compromiser interacts with a steadfast-type she will cave, thereby avoiding conflict. We have once again stumbled upon a hawk-dove game.

What does interpreting Locke's state of nature as a hawk-dove game buy us, exactly? If we consider the polymorphic equilibrium in which hawks and doves coexist, it seems to not get us too much – lots of conflict is still possible, meaning the state of nature is still a rather toxic place and agents have incentive to patronize protection agencies. Yet recent work on social networks and evolutionary game theory give us reason to be more optimistic. Santos and Pacheco (2005) have

uncovered that dovish becomes the norm when agents are embedded on a so-called scale-free network. [89] This is of particular interest because many real-world networks are considered to be scale-free, including some social networks.[90]

High levels of doves are possible in other circumstances as well. For example, if individuals are able to condition their behavior on the history of their opponent, then conflict can be kept to a minimum. As we have seen in the previous section, if agents are allowed to use tit-for-tat and are embedded on a lattice then hawks are rapidly expelled from the population. Additionally, we could imagine that the information-intensive strategies considered by Kavka such as Avenger and Guardian would work in such a setting as well, if the informational assumptions hold, that is. High levels of correlation would be another means of ensuring that conflict is abated.[91] Providing an exhaustive list of social mechanisms that promote compromise is beside the point, for all I desire to establish is that, contra Locke and Nozick, high levels of conflict are by no means guaranteed in a decentralized community of agents. Disagreement in the state of nature is an unavoidable fact, but violence is by no means inevitable. This presumably casts doubt on Nozick's origin story, which is predicated on the state of nature being a rather brutish place.[92] If conflict is highly unlikely, then agents may not feel the need to subscribe to a protection agency.[93]

---

[89] A scale-free network is a small-worlds network that is characterized by unequal connectivity, short path lengths and a high clustering coefficient. That is to say, some agents have significantly many more connections than others, any two agents randomly selected are only a few hops away from each other, and that it is very likely that those I am connected to are in fact connected to each other as well. As mentioned, these properties are often reflected in real-world networks.

[90] For example, networks of sexual contact have been shown to be scale free (Liljeros et al., 2001), as have networks of scientific communities and collaboration (Newman, 2004).

[91] By this we mean that in a mixed population agents playing the same strategy are more likely than chance to interact with one another. See Skyrms, 1996 for a demonstration of the effect correlation has in the prisoner's dilemma.

[92] Note that there have been many criticisms over the years of Nozick's invisible hand story. However, the vast majority of them have taken issue with either how Nozick dealt with competition between firms and the emergence of a dominant protection agency, or the existence of so-called "independents" who do not patronize a protection firm. I contend, however, that the story is suspect because firms may have great difficulty acquiring any customers to begin with.

[93] Of course while Nozick assumes the majority of individuals in the state of nature are good Lockeans that respect each other's rights, there could be a few "bad apples." Vanderschraaf (2006) shows that under certain assumptions this can result in wide-spread conflict, ultimately returning to the state of Hobbesian anticipation. Thus there may be impetus for a protection agency even if the vast majority of agents are Lockean doves.

4.5 Discussion

We have seen that, contra Kavka, a society composed of tolerant agents need not be unstable. Furthermore, unlike Kavka, our results do not require that we assume high levels of common knowledge. Agents in the society simply condition their actions on the past observed behavior of their counterparts. Such a simple mechanism, when agents are situated on a network, can generate high levels of compromise.

Chapter 5

CONCLUSION

We can now provide a partial answer to the question which initially motivated this dissertation. Cooperation, social contract formation and the equal split are possible even when we incorporate diversity into our theoretical models. Yet much depends on the details. As we saw in chapter 2, if traits are flexible, then social contract formation is nearly inevitable. Yet if traits are difficult to update, or even just get updated less frequently than the tolerance level, then collective action is difficult, if not impossible, to establish. Chapter 4 further drives this point home. Behavior in the Kavka-conflict game changes dramatically with the addition of social structure and history-dependent strategies such as tit-for-tat. Thus there is a certain sense in which the results of this dissertation echo the answer initially given in chapter 1: cooperation is possible in certain circumstances, and unattainable in others. In the course of this dissertation we have identified some social mechanisms that promote pro-social behavior in the face of diversity, but this is by no means an exhaustive list, and further study is required to uncover a fuller picture.

There are a number of ways to naturally extend our study. First and foremost, the results of chapter 2 on the evolution of the stag hunt could be replicated on a lattice. Embedding the model in this slightly more realistic setting may have very interesting implications. In particular, we would expect to see geographic differences in traits, as well as geographic differences in tolerance levels. While there could exist certain regions of the lattice of high tolerance, there may be more parochial regions, in which while all are hunting stag, tolerance levels are low and the region is characterized by lots of homogeneity. This result would be desirable because it captures a phenomena we see in the world (i.e., segregation and high levels of intolerance). Additionally, placing the model on a social network is interesting for it enables us to compare different networks, thereby allowing us to better understand what network topologies generate widespread tolerance and cooperation.

The results of chapter 3 can be extended in a natural way as well. Specifically, while we studied a collective action game and a resource division game separately, we could easily combine these to get a more holistic picture of how a social contract would evolve in the presence of fixed

markers. Wagner (2012) investigates the combined stag hunt and Nash division game, but does not incorporate cheap talk. What he finds is that the combined game enlarges the basin of attraction for both the stag hunting equilibrium as well as the equal split equilibrium. This is encouraging, and my best guess would be that both collective action as well as the equal split is all the more likely when agents are allowed to utilize plastic signals. I am less confident, however, that incorporating fixed signals into this model will promote egalitarian distributions and stag hunting.

There are other mechanisms not explored in the course of this dissertation that can minimize conflict in a divided community. Consider, for example, a modus vivendi. In this case two groups make an agreement or compromise that temporarily guarantees peace. Many consider a modus vivendi as unsatisfactory, because such an agreement is taken to be exceptionally unstable. Rawls, for instance, argues that a modus vivendi is inherently unstable because the agreement will no longer be respected once one group seizes power. In other words, the fortuitous group will immediately disregard the modus vivendi and oppress rival groups ("Social unity is only apparent, as its stability is contingent on circumstances remaining such as not to upset the fortunate convergence of interests." Rawls, 1993). Peace is fleeting. Yet I think this is much too fast, and I'd like to briefly argue that a modus vivendi may be much more stable than Rawls appreciates.

Consider a community with two groups, A and B. In each period exactly one group is in power, and the group in power can either "oppress" the rival group or "let live." Groups in power get a higher payoff in the current round if they oppress, and not surprisingly, the out-of-power group prefers they not be oppressed. We now introduce some uncertainty in order to make this a stochastic game. With some probability, call it p, the current group in power secures their privileged position in the next round. With probability 1-p, the ruling group falls out of power and the rival group takes the reigns in the next period. Groups can now condition their behavior on the past behavior of their rival. Consider, for example, the strategy: "if in power, oppress if counterpart has oppressed in the past, let live if else." Although I will refrain from going into much technical detail, it is easy to see that two groups utilizing the above mentioned strategy will, under many parameter settings, not desire to unilaterally deviate. Further work must be done to assess how stable a modus vivendi is in more realistic settings, where, for example, the chance of

transitioning from regimes (i.e., p) is endogenous. A full discussion of this is outside the bounds of this dissertation, and will be pursued at a later date. Nonetheless, it is clear that the contents of this dissertation are only the first steps toward a more complete understanding of conflict in a diverse society.

BIBLIOGRAPHY

Alexander JM (2007) *The Structural Evolution of Morality*. Cambridge University Press.

Antal T, Ohtsuki H, Wakeley J, Taylor P and Nowak M (2009) Evolution of cooperation by phenotypic similarity. *Proc. Natl. Acad. Sci. USA* 106: 8597-8600.

Aumann, R. (1974) Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1: 67-96.

Axelrod R (1984) *The Evolution of Cooperation*. NY: Basic Books.

Axelrod R, Hammond RA and Grafen A (2004) Altruism via kin-selection strategies that rely on arbitrary tags with which they coevolve. *Evolution* 58: 1833-1838.

Axelrod R, R Hammond (2006) The evolution of ethnocentrism. *The Journal of Conflict Resolution* 50: 926-936.

Axtell R, Epstein J and Young P (2006) The emergence of class in a multi-agent bargaining model. In *Generative Social Sciences: Studies in Agent-Based Computational Modeling*, ed. Joshua Epstein. Princeton: Princeton University Press.

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

Binmore K (2005) *Natural Justice*. Oxford University Press.

Binmore, K. (1998) *Game Theory and the Social Contract: Just Playing.* Cambridge: MIT Press.

Boyd, R., Gintis, H., Bowles, S. and P. Richerson (2003) The evolution of altruistic punishment. PNAS, 100: 3531-3535.

Braithwaite, R. (1955) Theory of Games as a tool for the Moral Philosopher. Cambridge University Press.

Bruner J (forthcoming) Diversity, tolerance and the social contract. Politics, Philosophy and Economics.

Cave, E. (1996) Would pluralist angels (really) need government? Philosophical Studies, 81: 227-246.

Dawkins R (1987) *The Extended Phenotype*. Oxford University Press.

Frederickson, G. (2002) *Racism: A Short History.* Princeton, NJ: Princeton University Press.

Friedman, M (1962) *Capitalism and Freedom*. The University of Chicago Press.
Gallo, E. (manuscript) Bargaining and social structure.

Gaus, G. (2011) *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge University Press.

Gauthier D (1969) *The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes*. Oxford University Press.

Glaeser E, Laibson D, Scheinkman J and Soutter C (2000) Measuring trust. *The Quarterly Journal of Economics* 155: 811-846.

Grim P, Selinger E, Braynen W, Rosenberger R, Au R, Louie N, J Connolly (2005) Modeling prejudice reduction: spatialized game theory and the contact hypothesis. *Public Affairs Quarterly* 19: 95-125.

Grim, P., Au, R., Louie, R., Rosenberger, R., Braynen, W., Selinger, E., and R. Eason (2008). A graphic measure of game-theoretic robustness, *Synthese*, 163: 273-297.

Grimm, P. (1995) The greater generosity of the spatialized prisoner's dilemma. Journal of Theoretical Biology, 173: 353-359.

Hales D (2005) Change your tags fast! A necessary condition for cooperation? *Multi-Agent and Multi-Agent Based Simulations* 3415: 89-98.

Hamilton WD (1964) The genetical evolution of social behavior II. *Journal of Theoretical Biology* 7:17-52.

Hammond, R. and R. Axelrod (2006) The evolution of ethnocentrism. Journal of Conflict Studies, 50: 926-936.

Hauert, C. and M. Doebli (2004) Spatial structure often inhibits the evolution of cooperation in the snowdrift game. Nature, 428: 643-6.

Henry, P. and D. Sears (2002) The symbolic racism 2000 scale. Political Psychology, 23: 253-283.

Hobbes T (1994) *Leviathan.* Ed. E Curley, Indianapolis: Hackett Publishing Co.

Hume D (2000) *A Treatise of Human Nature*. Oxford University Press, eds. D F Norton and M. J. Norton edition.

Huttegger S and Zollman K (2013) Methodology in biological game theory. *The British Journal for the Philosophy of Science*, 64: 637-658.

James A (2012) *Fairness in Practice: A social contract for a global economy.* Oxford University Press.

Jansen V and Baalen M (2006) Altruism through beard chromodynamics. *Nature* 440: 663-666.

Kandori, M. (1992) Social norms and community enforcement. *Review of Economic Studies*, 59: 63-80.

Kavka G (1995) Why even morally perfect people would need government. *Social Philosophy and Policy* 12: 1-18.

Kavka G (1986) *Hobbesian Moral and Political Theory*. Princeton University Press.

Krupp D, Debruine L, and Barclay P (2007) A cue of kinship promotes cooperation for the public good. *Evolution and Human Behavior* 29: 49-55.

Liljeros, F., Edling, C., Amaral, L., Stanley, H. and Y. Aberg (2001) The web of human sexual contacts. Nature, 411: 907-8.

Locke, J. (1988). *Two Treatises of Government*, ed. Peter Laslett. Cambridge University Press.

Locke, J. (2010) *An Essay Concerning Tolerance: And other writings on law and politics, 1667-1683*. Ed. Milton, J. and P. Milton. Oxford University press.

Mallon, R. (2006) Race: normative, not metaphysical or semantic. *Ethics*, 116: 525-551.

Mallon, R. (2007) A field guide to social construction, *Philosophy Compass*, 2: 93-108.

Miguel E and Gugerty M (2005) Ethnic diversity, social sanctions and public goods in Kenya. *Journal of Public Economics* 89: 2325-2368.

Moehler M (2009) Why Hobbes' state of nature is best modeled by an assurance game. *Utilitas* 21: 297-326.

Muldoon R, Borgida M, M Cuffaro (2012) The conditions of tolerance. *Politics, Philosophy and Economics* 11: 322-344.

Newman, M. (2004) Coauthorship networks and patterns of scientific collaboration. PNAS, 101: 5200-5.

Nowak M (2006) Five rules for the evolution of cooperation. *Science*, 314: 1560-1563.

Nowak, M. and R. May (1992) Evolutionary games and spatial chaos. Nature, 359: 826-9.

Nozick, R. (1974) *Anarchy, State and Utopia*.

Pollack G (1989) Evolutionary stability of reciprocity in a viscous lattice. *Social Networks* 3: 175-212.

Queller D, Ponte E, Bozzaro S and Strassmann J (2003) Single-gene greenbeard effects in the social amoeba dictyostelium discoideum. *Science* 299: 105-106.

Rawls J (1971) *A Theory of Justice*. Cambridge: Harvard University Press.

Rawls J (1993) *Political Liberalism*. New York: Columbia University Press.

Riolo R, Cohen M and Axelrod R (2001) The evolution of cooperation without reciprocity. *Nature* 414: 441-443.

Roberts G and Sherratt T (2002) Does similarity breed cooperation? *Nature* 418: 499-500.

Robson, A. (1990) Efficiency in evolutionary games: Darwin, Nash and the secret handshake. *Journal of Theoretical Biology*, 144: 379-396.

Rousseau J (1755) *A Discourse on Inequality*. Penguin Books, Trans. M. Cranston (1984) edition.

Sandholm, W. (2010) *Population Games and Evolutionary Dynamics*. MIT Press.

Santos, F and J. Pacheco (2005) Scale-free networks provide a unifying framework for the emergence of cooperation. Physical Review Letters, 95: 098104.

Schelling, T. (1978) *Micromotives and Macrobehavior*. New York: Norton.

Simmons, J. (1989) Locke's state of nature. *Political Theory*, 3: 449-70.

Simmons, J. (1993) *On the Edge of Anarchy: Locke, Consent and the Limits of Society*. Princeton University Press.

Singer P (1981) *The Expanding Circle: Ethics and Sociobiology*. Princeton University Press.

Skyrms (1996) *The Evolution of the Social Contract*. Cambridge University Press.

Skyrms (2004) *The Stag Hunt and the Evolution of Social Structure.* Cambridge University Press.

Skyrms B (2002) Signals, evolution and the explanatory power of transient information. *Philosophy of Science* 69: 407-428.

Skyrms B (2013) Natural social contracts. *Biological Theory* 8: 179-184.

Skyrms B and Zollman K (2010) Evolutionary considerations in the framing of social norms. *Politics, Philosophy & Economics* 9: 265-273.

Smead R and Forber P (forthcoming) An evolutionary paradox for prosocial behavior. *Journal of Philosophy*.

Smead R and Huttegger S (2011) Efficient social contracts and group selection. *Biology and Philosophy*, 26: 517-531.

Smith, J and E Szathmary (2000) *The Origins of Life: From the Birth of Life to the Origin of Language*. Oxford University Press.

Sugden, R. (1986) *The Economics of Rights, Co-operation and Welfare*. Oxford: Basil Blackwell, Inc.

Vanderschraaf, P (2006) War or peace?: A dynamical analysis of anarchy. Economics and Philosophy, 22: 234-279.

Vanderschraaf, P (2007) Covenants and reputations. Synthese, 157:167-195.

Wagner (2012) Evolving to divide the fruits of cooperation. *Philosophy of Science*, 79: 81-94.

Wagner, E. (manuscript) The long run stability of collective action.

Weibull J (1995) *Evolutionary Game Theory*. Cambridge: MIT Press.

Young, P. (1996) The economics of convention. *The Journal of Economic Perspectives*, 10: 105-122.

Young, P. (1993) The evolution of conventions. *Econometrica*, 61: 57-84.

Young, P. and D. Foster (1991) Cooperation in the short and in the long run. *Games and Economic Behavior*, 3: 145-156.

Zollman (2005) Talking to Neighbors: the evolution of regional meaning. *Philosophy of Science*, 72: 69-85.