A Cooperative Species:
Human Reciprocity and its Evolution
Contents

Preface x

1 A Cooperative Species 3

2 The Evolution of Altruism in Humans 12
  2.1 Preferences, Beliefs, and Constraints 13
  2.2 Social Preferences and Altruistic Cooperation 15
  2.3 Genes, Cultures, Groups, and Institutions 19
  2.4 Preview 25

3 The Biology of Self-Interested Cooperation 28
  3.1 Kin Altruism 30
  3.2 Reciprocal Altruism 33
  3.3 Reciprocal Altruism in Large Groups 38
  3.4 Reputation: Indirect Reciprocity 44
  3.5 Altruism as a Signal of Quality 48
  3.6 The Limits of Mutualistic Cooperation 49

4 Cooperative Homo economicus 51
  4.1 Folk Theorems and Evolutionary Dynamics 52
  4.2 The Folk Theorem with Imperfect Public Information 56
  4.3 The Folk Theorem with Private Information 60
  4.4 Dynamically Irrelevant Equilibria 62
  4.5 Social Norms and Correlated Equilibria 64
  4.6 The Missing Choreographer 67

5 Social Preferences 70
  5.1 The Structure of Human Morality 70
  5.2 Preferences in the Laboratory 71
  5.3 Strong Reciprocity is Common 76
  5.4 Free-riders Cause Cooperation to Unravel 78
## Contents

5.5  Altruistic Punishment Sustains Cooperation 81  
5.6  Effective Punishment Depends on Legitimacy 83  
5.7  Purely Symbolic Punishment is Effective 88  
5.8  People Punish Those Who Hurt Others 89  
5.9  Social Preferences are not Irrational 90  
5.10 Cultures and Institutions Matter 92  
5.11 Behavior is Conditioned on Group Membership 95  
5.12 People Enjoy Cooperating and Punishing Free-riders 98  
5.13 Competing Explanations 100

6  Multi-level Selection and the Evolution of Altruism 105  
6.1 Multi-level Selection 108  
6.2 Equilibrium Selection 113  
6.3 Positive Assortment 116  
6.4 Models and Motives 120

7  Ancestral Human Society 123  
7.1 Cosmopolitan Ancestors 126  
7.2 Genetic Evidence 131  
7.3 Prehistoric Warfare 136  
7.4 The Foundations of Social Order 142  
7.5 The Crucible of Cooperation 147

8  The Coevolution of Institutions and Behaviors 148  
8.1 Within-group Variance Reduction 150  
8.2 Selective Extinction 152  
8.3 Reproductive Leveling 155  
8.4 Empirical estimates of $F_{ST}$ 158  
8.5 Deme Extinction and the Evolution of Altruism 160  
8.6 The Australian Laboratory 161  
8.7 The Coevolution of Institutions and Altruism 163  
8.8 Simulating Gene-culture Coevolution 166  
8.9 Levelling, Competition, and Cooperation 170

9  Parochialism Altruism and War 173  
9.1 Parochial Altruism and War 176  
9.2 Results 179  
9.3 Persistent Parochialism 185
9.4 The Legacy of ‘Red in Tooth and Claw’ 187

10 The Evolution of Strong Reciprocity 190
  10.1 Punishment with Quorum Sensing 192
  10.2 Altruistic Punishment in a Realistic Demography 200
  10.3 Simulation Results 203
  10.4 Why Quorum-based Punishment Succeeds 209
  10.5 Retribution and Social Order 210

11 Socialization 213
  11.1 Cultural Transmission and Socialization 215
  11.2 When is Virtue its Own Reward? 218
  11.3 The Coevolution of Genes and Culture in Norm Internalization 221
  11.4 Altruism as Hitchhiker 225
  11.5 How Altruism Survives Payoff-based Updating 228
  11.6 Why is Internalization Socially Beneficial? 230
  11.7 The Programmable Brain 235

12 Social Emotions 237
  12.1 Reciprocity, Shame, and Punishment 240
  12.2 The Evolution of Social Emotions 243
  12.3 The Great Captains of our Lives 245

13 Conclusion: Human Cooperation and its Evolution 248
  13.1 The Origins of Human Cooperation 249
  13.2 The Future of Cooperation 253

14 Appendices 256
  A1 Agent-Based Models 256
  A2 Game Theory 261
  A3 Continuation probability and Time Discount Factor 265
  A4 Alternatives to the Standing Model 266
  A5 The Prisoner’s Dilemma with Public and Private Signals 267
  A6 Student and Nonstudent Experimental Subjects 270
  A7 Dyamical Systems 272
  A8 Altruism Defined 275
  A9 The Price Equation 277
Contents

A10 Weak Multi-level Selection 282
A11 Cooperation and Punishment with Quorum Sensing 283
A12 Social Emotions 292

Table of Symbols 294

References 296

Index 333
Preface


We have presented technical material in verbal as well as mathematical form wherever possible, and avoided mathematical formulations entirely where that is possible without sacrificing clarity. The appendices include brief description of several techniques that we employ throughout the book. When a technical term is first used, the term appears in italics and is defined in the Index. The page number in which a definition will be found is also in italics. Further uses of technical terms refer back to the definition, or to an explanation in the Appendices, using the symbol § followed by chapter number (if not in the current chapter) and section number. More complete statements of some of the technical aspects of our models and simulations are found in the works just cited. Readers interested in the overall argument of the book may wish to read chapters 1, 2, and 13. Sections §5.3 to §5.12, on the experimental evidence concerning social preferences, and sections §4.1 to §4.4, on economic models of cooperation based on repeated game theory, may be skipped without losing the main themes of the book.

We are grateful to the South African Museum, Cape Town for permission to use the Zaamenkamst Panel from Maclear district in the southern Drakensberg Mountains. [Permission to be requested] Thanks also to Margaret Alexander, Joy Lecuyer, Timothy Taylor, and Della Ulibarri at the Santa Fe Institute, and the hospitality of the Certosa di Pontignano and the Scuola Superiore Santa Chiara of the University of Siena. Finally, we gratefully acknowledge the John D. and Catherine T. MacArthur Foundation, University of Siena, Central European University, the Hungarian Scientific Research
Fund (OTKA), the U.S. National Science Foundation, and especially the Behavioral Science Program at the Santa Fe Institute, for providing research support.

Santa Fe, NM USA
Northampton, MA USA
September 2010
1

A Cooperative Species

How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortunes of others, and render their happiness necessary to him, though he derives nothing from it, except the pleasure of seeing it.


Is our conscience nothing but “the inner voice that tells us that somebody might be looking,” as the jaundiced H. L. Mencken (1949) put it? Or did the 20th century American essayist overlook humanity’s penchant genuinely to care for others, including total strangers, and to act morally, even when nobody is looking. But if Adam Smith is more right than Mencken, how could this oddly cooperative animal, *Homo sapiens*, ever have come to be?

In the pages that follow we advance two propositions. First, people cooperate not only for self-interested reasons but because they are genuinely concerned about the well being of others, try to uphold social norms, and value behaving ethically for its own sake. People punish those who exploit the cooperative behavior of others for the same reasons. Contributing to the success of a joint project for the benefit of one’s group, even at a personal cost, evokes feelings of satisfaction, pride, even elation. Failing to do so is often a source of shame or guilt. Second, we came to have these “moral sentiments” because our ancestors lived in environments, both natural and socially constructed, in which groups of individuals who are predisposed to cooperate and uphold ethical norms tended to survive and expand relative to other groups, thereby proliferating these pro-social motivations. The first proposition concerns proximate motivations for prosocial behavior, the second addresses the distant evolutionary origins and ongoing perpetuation of these cooperative dispositions.

Cooperation was prominent among the suite of behaviors that marked the emergence of behaviorally modern humans in Africa. Those living 75-90,000 years ago at the mouth of what is now the Klasies River near
Port Elizabeth, South Africa, for example, consumed eland, hippopotamus, and other large game. Among the remains found there is a now-extinct giant buffalo *Pelovoris antiquus* that weighed almost 2000 kilograms and whose modern day (smaller) descendant is one of the most dangerous game animals in Africa (Milo 1998). The Klasies River inhabitants, and their contemporaries in other parts of Africa, almost certainly cooperated in the hunt and shared the prey among the members of their group. Evidence of even earlier trade in exotic obsidians extending over 300 kilometers in East Africa, provides another unmistakable footprint of early human cooperation.

Like those living at Klasies River mouth, other ‘hunting apes’ quite likely cooperated in the common projects of pursuing large game, sharing the prey and maintaining group defense. Both *Homo neanderthalensis* and the recently discovered *Homo floresiensis* survived well into the late Pleistocene (about 126,000 to 11,000 years before the present) and hunted large game, the latter targeting the pygmy (but nonetheless substantial) elephants that had evolved on the island environment of Flores, off the coast of Indonesia.

Other primates engage in common projects. Chimpanzees, for example, join common defensive patrols and some hunt cooperatively. Male Hamadyras baboons respect property rights in food and mates. Many species breed cooperatively, with helpers and baby sitters devoting substantial energetic costs to the feeding, protection and other care of non-kin (Hrdy 2009). Social insects, including many species of bees and termites, maintain high levels of cooperation, often among very large numbers of individuals. Other common forms of cooperation among non-human animals, summarized by Kappeler and van Schaik (2006) are “grooming and other forms of body care, alarm calling, predator inspection, protection against attacks by predators or conspecifics, supporting injured group members...[and] egg-trading among hermaphrodites.”

Cooperation is common in many species. But cooperation in *Homo sapiens* is exceptional in that human cooperation extends beyond close genealogical kin to include even total strangers, and occurs on a much larger scale than other species except for the social insects.

In the pages that follow we will examine the cultural, biological and other processes that explain how humans became this exceptionally cooperative species.

By cooperation we mean engaging with others in a mutually beneficial activity. Examples include the joint pursuit of political and military objec-
Cooperative behavior may confer benefits net of costs on the individual cooperator, and thus may be motivated entirely by self-interest. In this case, cooperation is a form of mutualism, namely an activity that confers net benefits both on the actor and on others. But, cooperation may also impose net costs upon the individual. In this case cooperative behavior constitutes a form of altruism. By contrast to mutualistic cooperation, altruistic cooperation would not be undertaken by an individual whose motives were entirely self-interested and thus did not take account of the effects of one’s actions on others.

The evolution of cooperation that is mutualistic or involving only close family relatives is easily explained. Cooperation among close family members could have evolved by natural selection because the benefits of cooperative actions are conferred on the close genetic relatives of the cooperator, thereby helping to proliferate alleles ('genes') associated with the cooperative behavior. Cooperation could also have evolved because one individual’s costly contribution to the welfare of another individual is reliably reciprocated at a future date, thereby making cooperation mutualistic. Models of altruism towards close family members and reciprocal altruism, which really should be called “enlightened self-interest,” are popular among biologists and economists alike and explain many forms of human cooperation, particularly those occurring among close kin or in dyadic (two-person) or other very small group interactions.

But these models fail to explain two facts about human cooperation: that it takes place in groups far larger than the immediate family, and that both in real life and in laboratory experiments, it occurs in interactions that are unlikely to be repeated, and where it is impossible to obtain reputational gains from behaving prosocially.

The most parsimonious proximal explanation of cooperation, supported by extensive experimental and other evidence, is that people enjoy cooperating, or feel morally obligated to cooperate, with like-minded people. People also enjoy punishing those who exploit the cooperation of others, or feel morally obligated to do so. Free-riders frequently feel guilt, and if they are sanctioned by others, they may feel ashamed. We term these feelings social preferences. Social preferences include a concern for the well being of others (positive or negative) and a desire to uphold ethical norms.
In many human groups, these motives are sufficiently common to sustain socially valuable norms that support contributions to projects of common benefit, even when cooperators bear costs in order to benefit others. The forms of cooperation and the behaviors that elicit punishment by peers differ from society to society, but the critical role of social preferences in sustaining altruistic cooperation is ubiquitous.

Because we are convinced that people enjoy cooperating and dislike people who do not, the task we will set is not that typically addressed by biologists and economists, namely to explain why people cooperate despite being selfish. Rather, we seek to explain why we are not purely selfish—why the social preferences that sustain altruistic cooperation are so common. Why do so many people care about fairness and reciprocity and value the well-being of fellow members of their groups, often favoring them over outsiders? Proximate answers to this question are to be found in the way that our brains process information and induce the behavioral responses that we term cooperation. But how did we come to have brains that function in this manner? Early human environments are part of the answer.

Early modern humans inhabited the large, mammal-rich African savannah and other environments in which cooperation in acquiring and sharing food yielded substantial benefits at relatively low cost. The slow human life-history with prolonged periods of dependency of the young also made the cooperation of non-kin in child rearing beneficial. As a result, members of groups that sustained cooperative strategies for provisioning, child-rearing, sanctioning non-cooperators, defending against hostile neighbors, and sharing truthfully transmitted information had significant advantages over members of non-cooperative groups.

In the course of our subsequent history we created novel social and physical environments exhibiting similar, or even greater, benefits of cooperation, among them the division of labor coordinated by generalized exchange and respect of rights of property, systems of production characterized by increasing returns to scale (irrigated agriculture, modern industry, information systems with network externalities), and warfare. The impressive scope of these modern forms of cooperation was facilitated by the emergence in the last seven millennia of governments capable of enforcing property rights and providing incentives for the self-interested to contribute to common projects.

Both prior to the emergence of governments and since, however, cooperation has been sustained also by motives that led some people to bear costs on
behalf of others, contributing to common projects, punishing transgressors, and excluding outsiders. These altruistic social preferences supporting cooperation out-competed unmitigated self-interest and proliferated for three reasons.

First, human groups have devised ways to protect their altruistic members from exploitation by the self-interested. Prominent among these is the public spirited shunning, ostracism, and even execution of free-riders and others who violate cooperative norms. Other group activities protecting altruists from exploitation are levelling practices that limit hierarchy and inequality, including the sharing of food and information.

Second, humans adopted prolonged and elaborate systems of socialization that lead individuals to internalize the norms that induce cooperation, so that contributing to common projects and punishing defectors became objectives in their own right rather than constraints on behavior. Together the internalization of norms and the protection of the altruists from exploitation were sometimes sufficient to halt entirely or even reverse within-group selection pressures operating against those who were motivated to bear personal costs to benefit others.

Third, between-group competition for resources and survival was a decisive force in human evolutionary dynamics. Groups with many cooperative members tended to survive these challenges and to occupy the territory of the less cooperative groups, thereby both gaining reproductive advantages and proliferating cooperative behaviors through cultural transmission. From warfare and environmental catastrophe among hunter-gatherers to the rise and fall of modern nation states, group extinction, costly group dispersal, and ostracism from groups have been powerful mechanisms supporting the evolution of human cooperation. The extraordinarily high evolutionary stakes of intergroup competition and the contribution of altruistic cooperators to success in these contests meant that sacrifice on behalf of others, extending beyond the immediate family and even to virtual strangers, could proliferate.

This is part of the reason why humans became extraordinarily group-minded, favoring cooperation with insiders and expressing hostility towards outsiders. Boundary-maintenance sustained within-group cooperation and exchange by limiting group size and within-group linguistic, normative and other forms of heterogeneity while at the same time sustaining the between-group conflicts and differences in behavior that make group competition a powerful evolutionary force.
In short, humans became a cooperative species because cooperation was highly beneficial to the members of groups that practiced it, and we were able to construct social institutions, to enforce norms, to share food, to socialize new members, to distinguish insiders from outsiders, to make war, all of which minimized the within-group selective pressures operating against those with social preferences, while heightening the group-level advantages associated with the high levels of cooperation that these social preferences allowed. Adherence to these institutions across generations was secured through the cultural transmission of the values and beliefs that favored conformity to existing norms. These institutions proliferated because cooperation enhanced the chances that a group would survive as a biological and cultural entity in the face of environmental, military and other challenges.

Early humans were not alone in occupying territory and a feeding niche that made cooperation among group members highly advantageous. Indeed our ancestors competed with lions, hyenas, wild dogs and other possibly hominid cooperative hunters for the very same ungulates and other large mammals. Nor were our ancestors exceptional in the kinds of group competition for territory and other valued resources that made cooperation so essential to survival. Chimpanzees, too, engage in lethal contests between troops where winners gain territory and reproductive advantages. The same is true of species as diverse as meerkats and fire ants. Nor are humans exceptional in constructing their own physical and social environments. Beavers build dams, birds build nests, and burrowing animals build underground catacombs. Why then did humans, rather than chimps, lions, or meerkats, develop such exceptional forms of cooperation?

Central to our reply are the human cognitive, linguistic and physical capacities that made us especially good at all of the above, and more. These capacities allow us to formulate general norms of social conduct, to erect social institutions regulating this conduct, to communicate these rules and what they entail in particular situations, to alert others to their violation and to organize coalitions to punish the violators. No less important is the psychological capacity to internalize norms, to experience such social emotions as shame and moral outrage, and to base group membership on such non-kin characteristics as ethnicity and linguistic differences, which in turn facilitates costly conflicts among groups. Equally essential was the developmental plasticity of humans and our long period of maturation, the latter initially a result of the particular feeding niche that early humans occupied. Also important is the unique human capacity to use projectile weapons, a
consequence of which is to lower the cost of coordinated punishment of norm violators within a group, to reduce the costs of hunting large animals with concomitant benefits accruing to groups with widely endorsed sharing norms, to render intergroup conflicts more lethal, and hence to elevate group-level competition to a more powerful evolutionary force.

These exceptional aspects of human livelihoods and social interactions, we will show, have favored the evolution of an individual predisposition to cooperate with others and to punish those who exploit the cooperation of others. But more than individual-level motivation is involved. The regulation of social interactions by group-level norms and institutions plays no less a role than altruistic individual motives in understanding how this cooperative species came to be. Institutions affect the rewards and penalties associated with particular behaviors, often favoring the adoption of cooperative actions over others, so that even the self-regarding are often induced to act in the interest of the group. Of course it will not do to posit these rules and institutions *a priori*. Rather, we show that these could have co-evolved with other human traits in the relevant ancestral ecologies and social environments.

Had we chosen *A Colluding Species* as our title, it would not be necessary to point out that cooperation is not an end, but rather is a means. In some settings, competition, the antithesis of cooperation, is the more effective means to a given end. Similarly, the individual motives and group-level institutions that account for cooperation among humans include not only the most elevated, a concern for others, fair-mindedness, and democratic accountability of leaders, for example, but also the most wicked: vengeance, racism, religious bigotry, and hostility towards outsiders.

Price-fixing by cartels and other baleful economic effects of collusion motivated Adam Smith to advocate a competitive economic system under which such forms of anti-social collusion would unravel. In its stead he advocated “an invisible hand” that would guide the efforts of countless self-interested producers to coordinate a modern division of labor in the interest of all.

But if the late 18th century gave us this evocative metaphor for the beneficial effects of the pursuit of individual gain, the mid-20th century invented two no less riveting metaphors for the dark side of self-interest: the prisoner’s dilemma and the tragedy of the commons. Their logic inverted Adam Smith’s invisible hand, showing that even where cooperation was essential to the pursuit of common ends, it would falter in the face of self-interest.
Garrett Hardin’s tragedy of the commons was rapidly assimilated by scholars, as it embraced a model of self-interest already well established in both neoclassical economics and the neo-Darwinian synthesis in biology. Social preferences, Hardin made clear, were powerless to counter the “remorseless” degradation of the environment:

The tragedy cannot be solved by an appeal to conscience, for those who heeded the appeal would have fewer children, and by the heritability of capacity of conscience, this would lead to a less moral population. (p. 1246)

Because “freedom in a commons means ruin for all” (p. 1244),” he advocated a modern version of Thomas Hobbes’ Leviathan that he termed “mutual coercion mutually agreed upon” (p. 1247). Hardin termed his contribution a “rebuttal to the invisible hand” (p. 1244). Mancur Olson’s no less ineluctable “logic of collective action” in n-person prisoner’s dilemmas demonstrated the inevitability of a passive citizenry and the impossibility of cooperation, due to ubiquitous free riders.

But, as the prisoner’s dilemma and the tragedy of the commons were becoming staples of undergraduate instruction, field evidence from anthropologists and micro-historical studies of social movements pointed in an entirely different direction. High Alpine and Andean common summer pastures had averted tragedy without government regulation for centuries, possibly millennia. Workers and democrats had for centuries risked their lives in collective actions that plainly defied Olson’s logic. The work of Elinor Ostrom and her collaborators documented literally hundreds of decentralized tragedy-averting commons governance systems around the world, bringing to a head this collision of empirical observation and the logic of self-interest (Ostrom 1990).

The tension between the relentless logic of self-interest and the ubiquity of collective action in real world settings was eventually resolved by a series of experiments by psychologists and economists, most notably by Ernst Fehr and his colleagues (Fehr and Gächter 2000a, Herrmann et al. 2008). The experiments confirmed that self-interest is indeed a powerful motive, but also that other motives are no less important. Even when substantial sums of money are at stake, many, perhaps most, experimental subjects are fair-minded, generous towards those similarly inclined, and nasty towards those who violate these pro-social precepts. In light of these results, the evidence that the tragedy of the commons is sometimes averted and that
collective action is a motor of human history is considerably less puzzling. The puzzle, instead, is how humans came to be like this.


These recent works are reminiscent of Peter Kropotkin’s *Mutual Aid* a century earlier, a book that had advanced a kinder, gentler view of the evolutionary process in opposition to the then popular dog-eat-dog social Darwinist claims about what natural selection entails for human behavior. The moral, generous, and civic minded predispositions documented in these works and here in the pages that follow show that evolution can not only foster self-interest but also promote the generous and ethical behaviors that help us escape the prisoner’s dilemma and avert the tragedy of the commons, and permit us to sustain the hope for a society committed to freedom and justice for all. We will see that this is true not despite, but in important measure because, evolutionary processes are “red in tooth and claw,” in Alfred, Lord Tennyson’s famous words.
The Evolution of Altruism in Humans

The Americans... are fond of explaining almost all the actions of their lives by the principle of self-interest rightly understood, how an enlightened regard for themselves constantly prompts them to assist one another... In this respect I think they frequently fail to do themselves justice; in the United States as well as elsewhere people are sometimes seen to give way to those disinterested and spontaneous impulses that are natural to man; but the Americans seldom admit that they yield to emotions of this kind.


Like Alexis de Tocqueville’s Americans, a distinguished tradition in biology and the social sciences has sought to explain cooperative behavior “by the principle of self-interest, rightly understood.” Richard Dawkins (1989) in *The Selfish Gene*, writes “Let us try to teach generosity and altruism, because we are born selfish.” Similarly, drawing out the philosophical implications of the evolutionary analysis of human behavior, Richard Alexander (1987) writes, “ethics, morality, human conduct, and the human psyche are to be understood only if societies are seen as collections of individuals seeking their own self-interest.” (p. 3). From J. B. S. Haldane’s quip that he would risk his life to save eight drowning cousins (but not fewer) to the Folk Theorem of modern game theory (§4.1—§4.3), this tradition has clarified the ways that genetic relatedness, repeated interactions and reputation-building might confer fitness advantages and other benefits on those engaging in seemingly unselfish behaviors.

Our approach, however, favors Tocqueville, not Tocqueville’s ‘Americans.’ Explaining why will take us through disciplines as diverse as population genetics, evolutionary game theory, and archaeology and across semantic minefields of heavily freighted terms, such as altruism, and problems, such as the relationship between genetic inheritance and cultural transmission. In this chapter we explain our conceptual strategy and define terms.
Following William Hamilton, we use the term helping to describe behaviors that confer benefits on others, reserving the term altruism for helping in situations where the helper would benefit in fitness or other material ways by withholding help (a more complete definition is given in §A8). Our models and simulations in subsequent chapters show that these altruistic helping behaviors may proliferate under conditions under which ancestral humans lived.

### 2.1 Preferences, Beliefs, and Constraints

We explain the proximal influences on individual action by the beliefs, preferences, and constraints approach common to economics and decision theory. Restricted to a specific set of feasible actions, what individuals do in any situation depends on their desires and goals on the one hand, and their beliefs on the other. Constraints represent the limitations placed on the feasible actions an individual may take in a given situation. Beliefs are an individual’s representation of the causal structure of the world, including the relationship between choices and the relative probabilities of the various possible resulting outcomes. Preferences are the pro- or con-sentiments that make up the individual’s valuation of the various possible outcomes of making a choice.

Preferences may be described as an ordering (technically, a preference function) of the states of the world that may result from one’s actions. We assume preferences satisfy two conditions: they are complete (any two states can be compared) and transitive; that is, consistent, so that if one prefers A to B and B to C one then prefers A to C. Preferences are the results of a variety of influences: tastes (food likes and dislikes, for example), habits, emotions (such as shame or anger) and other visceral reactions (such as fear), the manner in which individuals construe situations (or more narrowly, the way they frame decisions), commitments (like promises), internalized norms of ethical behavior, psychological propensities (for aggression, extroversion and the like), and affective relationships with others.

As long as preferences are consistent, we can succinctly and analytically summarize the individual’s behavior as maximization of a preference function, even though this by no means describes the psychological processes underlying choice (Savage 1954). To say that individuals act on their preferences means that knowledge of these preferences provides a concise and
accurate account of their actions. Of course, this analytical account will not generally coincide with the account that individuals would give of their behavior.

The preferences, beliefs, and constraints approach is silent on the cognitive and other processes determining individual action. In some situations, buying a car, for example, individuals may deliberately optimize, while in others, diet or ethical behavior for example, they may follow rules of thumb that have been adopted without conscious optimization. Optimizing models are commonly used to describe behavior not because they mimic the cognitive processes of the actors, which they rarely do, but because they capture important influences on individual behavior in a succinct and analytically tractable way.

A version of the beliefs, preferences, and constraints model, incorporating the behavioral assumptions sometimes summarized as *Homo economicus*, has become standard throughout the human behavioral sciences. F. Y. Edgeworth, a founder of neoclassical economics, expressed this view in his *Mathematical Psychics* (Edgeworth 1881:104): “The first principle of economics is that every agent is actuated only by self-interest.” Self-interest is not presumed by rationality (rationality assumes individuals have transitive and complete preferences, although these preferences could be altruistic or even masochistic). Nevertheless, self-interest is commonly assumed in economics, and indeed sometimes even confused with rationality. Thus while self-interest is not formally implied by the conventional approach, it is generally assumed in practice. The assumption allows quite precise predictions in strategic situations when it takes the form of what we term the self-interest axiom namely self-interested individual preferences coupled with the belief that others are also motivated by self-interest.

But predictions based on Edgeworth’s self-interest axiom often fail to describe the actions people take. Indeed, the axiom was never intended to be taken literally. Edgeworth followed the statement above with the caveat that the axiom was strictly true only in “contract and war.” But, even in these areas, exceptions to the canon are glaring and increasingly well documented, as is shown by Truman Bewley’s (2000) finding that firms do not cut wages during recessions because wage cuts demoralize workers, who consider them unfair. Similarly, Jessica Stern’s (2003) finding that terrorist violence is motivated as a reaction against perceived injustice, and the case of kamikaze pilots (Hagoromo Society 1973), who volunteer to sacrifice
their lives out of a sense of honor and duty, are dramatic indications that people are often motivated by non-selfish principles.

The economist’s usual defense of the self-interest axiom is that it is self-evident, with the fallback assertion being that natural selection could not have produced any other kind of preferences. But, the axiom is far from self-evident and, as we will see in subsequent chapters, the assertion is simply false.

Examples of non-selfish motives include the importance of fairness considerations in wage setting and other exchanges (Blinder and Choi 2000). Equally at variance with self-interest is the fact that individuals bother to vote, given that the likelihood that their vote is decisive is vanishingly small, as well as the extensive support, when they do vote, for tax-financed income transfers to the poor even among those sufficiently rich and upwardly mobile to be very unlikely ever to benefit directly from these transfers (Gilens 1999, Fong 2001, Fong et al. 2005). Also telling against the self-evident status of the self-interest axiom are studies at Continental Airlines, Nucor Steel, and other companies that have found group incentives to be effective even where gain-sharing is distributed among such a large number that the additional income associated with one’s own effort is negligible (Hansen 1997, Knez and Simester 2001). Other examples include volunteering for dangerous military and other tasks, tax compliance far in excess of that which would maximize expected incomes (Andreoni et al. 1998), participating in various forms of collective action with little expectation of personal benefit (Moore, Jr. 1978, Wood 2003), and conforming to norms and laws in cases where one’s transgression would be personally advantageous and would not be detected.

2.2 Social Preferences and Altruistic Cooperation

Recall that social preferences are a concern for the well-being of others (positive or negative) and a desire to uphold ethical norms. By contrast with self-regarding preferences, which are based on states concerning oneself alone, we stress other-regarding preferences, which are valuations based at least in part on states that refer to others. Other-regarding preferences include not only generosity toward others and a preference for “fair” outcomes, but also what Thomas Hobbes called the desire for “eminence,” Thorstein Veblen’s “pecuniary emulation” exemplified by a desire to “keep up with the Joneses” (Veblen [1899]/1934), Charles Horton Coo-
ley’s “looking-glass self,” according to which our self-esteem is dependent in part upon what others think of us, so we attempt to favorably impress others as a means of raising our subjective self-esteem (Cooley 1902, Brennan and Pettit 2004), and Aristotle’s character virtues, such as honesty and courage, which are personal values that promote prosocial behavior (Aristotle 2002[350BC]).

Also important among the social preferences are ethical commitments, and these need not reflect a concern for the states experienced by others. One can be honest because one seeks to avoid imposing costs on others by deceiving them. But honesty could be entirely self-regarding, practiced in order to be the kind of person one wants to be. Other-regarding and ethical preferences may inflict costs on others when one wishes ill to others, and regards doing so to be one’s ethical duty. But generosity, fair-mindedness, and a commitment to honesty, for example, often motivate individuals to act in ways that benefit others. Because these aspects of social preferences are important in sustaining altruistic cooperation and because they are unconventional among economists and biologists, it is wise to clarify exactly what we mean.

We prefer the term “social preferences” to the more common but ambiguous “unselfish” or “non-self-interested” preferences. “Unselfish” behaviors are, like “selfish” behaviors, motivated by the individual’s preferences. If I get pleasure from helping others, or if I help others because I would feel guilty if I did not, I am no less motivated by my own preferences than if I enjoy eating a fine meal, or help another because I will be punished if I do not. Moreover, such other-regarding emotions as spite and envy, would not generally be termed “unselfish” in any sense. Nevertheless, like empathy, they are social preferences. The distinction between other-regarding and self-regarding preferences does not lie in other-regarding behaviors being counter-preferential, but rather in their being motivated at least in part by the effects of one’s actions on others.

We define a social dilemma as an interaction in which the actions of individuals result in an outcome that is socially inefficient in the (Pareto) sense that there exists some other feasible outcome such that at least one member could be better off while no member would be worse off. Examples of social dilemmas modeled by game theorists are the prisoner’s dilemma, the public goods game (§3.3), sometimes termed an n-person prisoner’s dilemma, the so-called “war of attrition” and other arms race interactions, the tragedy of the commons and the common pool resource game in which contributing
to the common project takes the form of forgoing the over-exploitation of a jointly-utilized resource such as a fishery, water supply, or forest. We say a person free rides if he benefits from the contributions of other group members while himself contributing less or nothing at all.

To clarify the distinction between self-regarding and other-regarding preferences, consider perhaps the most famous of all experimental games (§A2), the
\[
\text{prisoner's dilemma, with payoffs (the row player's first) shown in Figure 2.1. In this game, two subjects do not know each other's identity, will interact only once, and may not make any binding agreements about how they will play the game. This is an example of an anonymous, one-shot non-cooperative game (the latter term refers to the no binding agreements condition, not to the interests of the party or the outcomes of the game). The experimenter explains that each of the subjects can take one of two actions without knowing the action taken by the other: cooperate (C) or defect (D). If both choose to cooperate, each receives $10 (the intersection of the C row and the C column in the figure), and if both defect, each receives $5 (the intersection of the D row and the D column). Moreover, if one cooperates and the other defects, the defector gets $15 and the cooperator gets nothing (the off-diagonal payoffs in the figure).}
\]

\[
\begin{array}{cc}
C & D \\
C & 10,10 & 0,15 \\
D & 15,0 & 5,5 \\
\end{array}
\]

Figure 2.1. The Prisoners' Dilemma. Note: here and in other payoff matrices, the row player's payoff is first, and the column player's payoff is second.

Because a self-regarding player cares only about the payoff to himself, he reasons as follows. “If my partner cooperates, I get $15 by defect and $10 by cooperating, so I should defect. If my partner defects, I get $5 by defecting and nothing by cooperating, so I should still defect. Thus I should defect no matter what my partner does.” If both players are self-regarding, both will defect, and each will get $5, which is half of what they could have gotten by cooperating with each other. Thus, for a self-regarding person, Defect is a dominant strategy; that is, it is a best response (i.e., a payoff-maximizing strategy) irrespective of his beliefs about what the other will do. Because this is true for both Row and Column, mutual Defect is a dominant strategy.
strategy equilibrium. In general, a choice of strategies by players such that no player can gain from changing his strategy, no matter what the other players do.

An other-regarding player cares not only about his own payoff, but that of his partner as well. Such a player might reason as follows. “I feel sufficiently positive towards a partner who cooperates that I would rather cooperate even if by doing so I forgo the larger payoff ($15) I could have had by defecting. If my partner defects, I of course prefer to defect as well, both to increase my earnings, and to decrease the earnings of a person who has behaved uncharitably towards me.” If both players reason in this manner, and if each believes the other is likely to cooperate, both will cooperate. Thus, both mutual Cooperate and mutual Defect are Nash equilibria in a new game, transformed from the old by augmenting the material payoffs with the players concerns about one another. In general, we say that a choice of strategies by players is a Nash equilibrium. Note that a dominant strategy equilibrium is always a Nash equilibrium, but the reverse is not true. Social preferences may thus convert a prisoner’s dilemma material payoff structure into what is called an payoff structure—a player will cooperate if assured that the other will cooperate as well. Which of the two Nash equilibria will obtain depends on the players’ beliefs about what the other will do.

Despite the strong temptation to defect either out of selfishness or fear of being exploited by the other player, many experiments have found that a considerable fraction of subjects prefer to cooperate rather than defect in the prisoner’s dilemma (Sally 1995). One famous real-life, high stakes example is the popular TV show “Friend or Foe,” where contestants play a prisoner’s dilemma with stakes varying between $200 and $22,000. About half of contestants choose to cooperate even though they are guaranteed to earn more money by defecting, no matter what their partner does. Even more striking, contestants are no more likely to defect when the stakes are higher (Oberholzer-Gee et al. 2010, List 2006).

Similar behavior is observed in the laboratory. Kiyonari et al. (2000) had Japanese university students play the prisoner’s dilemma with real monetary payoffs. The experimenters ran three distinct treatments, with about equal numbers of subjects in each treatment. The first treatment was a standard simultaneous prisoner’s dilemma in which both players choose whether to cooperate or defect without knowing his partner’s choice. The second was a sequential “second player” prisoner’s dilemma in which one player had to
choose between cooperate and defect after being informed that his partner had already chosen to cooperate. The third was again a prisoner’s dilemma, which we will call a “first player” prisoner’s dilemma, in which a player was told that he would choose first, but his decision to cooperate or defect would be transmitted to his partner before the latter made his own choice. The experimenters found that 38% of the subjects cooperated in the standard simultaneous treatment, 62% cooperated in the second player treatment, and 59% cooperated in the first-player treatment. The decision to cooperate in each treatment cost the subject about $5 (600 yen). This shows unambiguously that a majority of subjects were conditional altruistic cooperators (62%). Almost as many were not only cooperators, but were also willing to bet that their partners would be (59%), provided the latter were assured of not being defected upon, although under standard conditions, without this assurance, only 38% would in fact cooperate. Experiments conducted by Hayashi et al. (1999), Watabe et al. (1996), Morris et al. (1998), McCabe et al. (2000) and Clark and Sefton (2001) found similar subject behavior.

Chapter 5 presents experimental and other evidence that social preferences are common in most cultures for which we have evidence. The same evidence shows that the fraction of most populations motivated solely by self-regarding preferences is quite modest.

2.3 Genes, Cultures, Groups, and Institutions

We define culture as the ensemble of shared preferences, knowledge, and beliefs that are acquired by means other than genetic transmission. Culture is an evolutionary force in its own right, not simply an effect of the interaction of genes and natural environments.

A more parsimonious approach holds instead that while predispositions that are transmitted culturally may constitute the proximate causes of behavior, they in turn are entirely explained by the interaction of human nature and the natural environment. According to this view, for example, the Lamalera whale hunters we discuss in Chapter 5 would be said to share valued resources because they have social preferences, but they have social preferences because they live in a place where hunting whales is the best way to make a living, and those who hunt large game do better if they learn how to share.

It is of course true that the interaction of natural environments and genes affects the evolution of cultures has yielded numerous insights. But, it is
also true that culture affects the natural and social environments in which the relative fitness of genetically transmitted behavioral traits is determined. Luigi Luca Cavalli-Sforza and Marcus Feldman (1981), Robert Boyd and Peter Richerson (1985), William Durham (1991), and Richerson and Boyd (2004) and others have provided compelling instances of these cultural effects on genetic evolution. It follows that human cognitive and affective attributes are the product of a dynamic whereby genes affect cultural evolution and culture affects genetic evolution, known as *gene-culture coevolution*. To see how gene-culture coevolution works, think about the ways that an organism may acquire information.

The genome encodes information that is used both to construct a new organism, to instruct the new organism how to transform sensory inputs into decision outputs (i.e., to endow the new organism with a specific preference structure), and to transmit this coded information virtually intact to the new organism. Since learning about one’s environment is costly and error-prone, efficient information transmission is likely to ensure that the genome encode information relevant to aspects of the organism’s environment that are constant, or that change only very slowly through time and space. By contrast, environmental conditions that vary across generations and/or in the course of the organism’s life history can be dealt with by providing the organism with the capacity to learn from one’s environment, and hence phenotypically adapt to specific conditions.

For most animals, genetic transmission and individual learning are about all there is as far as information acquisition is concerned. Humans, by contrast, also acquire information from one another through a process of social learning. To see just how inadequate individual learning and genetically-transmitted information would be in supporting human life, consider the four hapless Europeans who in 1860 attempted to cross the Australian continent from south to north and back, armed only with their ability to devise ways of living in an unfamiliar environment with the help of then-sophisticated equipment and ample stocks of food, carried on imported camels (Henrich and McElreath 2003). After a series of reverses, and having eaten the unfortunate camels, they resorted to foraging, attempting vainly to learn how to trap rats and birds and to catch fish in the occasional well-watered spot. Despite the generous gifts of food from the aboriginal groups who they encountered as they struggled on, three perished and the last was saved by a community of Yantruwanta people living in the desert, where he fully recovered and was eventually found by a European search
party. Even more striking is the fact that the Norse settlers mentioned in chapter 9, who lived in Greenland, over five centuries of trial and error, did not learn the elements of survival in that climate that were well-known to the Inuit living nearby (McGhee 1984). Thus there is a distinctively human intermediate case that is not well-handled by either genetic encoding or learning from one’s environment de novo in each generation. When environmental conditions are positively but imperfectly correlated across generations, each generation acquires valuable information through learning that it cannot transmit genetically to the succeeding generation, because such information is not encoded in the germ line. In such environments, an animal could benefit from the transmission of information concerning the current state of the environment through some non-genetic information channel. Such information, called epigenetic by biologists, is quite common (Jablonka and Lamb 1995) and achieves its highest and most flexible form in cultural transmission in humans and to a considerably lesser extent in other primates (Bonner 1984, Richerson and Boyd 1998). Cultural transmission, also called social learning as opposed to individual learning, takes the form of vertical (parents to children) horizontal (peer to peer), and oblique (non-parental elder to younger) transfer of information. Examples are given by Cavalli-Sforza and Feldman (1981) and Henrich and Gil-White (2001) who, find that higher status individuals influence the cultural values of lower status individuals, Newman et al. (2006), who find that individuals tend to differentially embrace highly popular cultural practices.

The parallel between cultural and biological evolution goes back to William James (1880) and Julian Huxley (1955). The idea of treating culture as a form of epigenetic transmission was pioneered by Cavalli-Sforza and Feldman (1973), Karl Popper (1979), and Richard Dawkins, who coined the term “meme” in The Selfish Gene (1976) to represent an integral unit of information that could be transmitted phenotypically. There quickly followed several major contributions to a biological approach to culture, all based on the notion that culture, like genes, could evolve through replication (intergenerational transmission), mutation, and selection (Lumsden and Wilson 1981, Cavalli-Sforza and Feldman 1982, Boyd and Richerson 1985).

Richard Dawkins added a second fundamental mechanism of epigenetic information transmission in The Extended Phenotype (1982), noting that organisms can directly transmit environmental artifacts to the next genera-
tion, in the form of such constructs as beaver dams, bee hives, and even social structures (e.g., mating and hunting practices). Creating an important aspect of its environment and stably transmitting this environment across generations, known as niche construction, is a widespread form of epigenetic transmission (Odling-Smee et al. 2003). Moreover, niche construction gives rise to what might be called a gene-environment coevolutionary process, since a genetically induced environmental regularity becomes the basis for genetic selection, and genetic mutations that give rise to mutant niches will tend to survive if they are fitness enhancing for their constructors.

Our own models of the co-evolution of genetically transmitted individual behaviors and culturally transmitted group-level institutions is another example of the same process. We will see (Chapter 7) that the presence of a culturally transmitted convention, resource sharing, is essential to the evolution of a genetically transmitted altruistic trait governed by natural selection. In Chapter 11 we show that the possibility of acquiring advantageous behaviors by social learning could generate the conditions under which a genetically transmitted capacity to internalize norms could evolve. Human cultures, along with the institutional structures they support, are instances of niche construction; i.e., the creation of a particular environment in such a manner that the genetic evolution of the creators is affected thereby (Laland et al. 2000, Bowles 2000, Laland and Feldman 2004).

In gene-culture coevolution, the evolutionary dynamic in group-structured populations accounts for human cooperation, including the proliferation of new behaviors as the result of learning from one’s own experiences and from others as well the transmission and selective replication of genetic and other information from parents to offspring. In this approach, the process of differential replication affects the frequency of both individual traits, generosity towards fellow group members, say, and group traits, a system consensus decision-making or property rights. Though inspired by biological approaches, especially Cavalli-Sforza and Feldman (1981), Boyd and Richerson (1985), and Durham (1991), like these authors, we do not privilege biological explanation. Our approach may be summarized as follows.

First, while genetic transmission of information plays a central role in our account, the genetics of social behavior is for the most part unknown. Knowledge of the genetic basis of the human cognitive and linguistic capacities that make cooperation on a human scale possible has expanded greatly
in recent years, but virtually nothing is known about genes that may be expressed in cooperative behavior, should these exist. No “gene for cooperation” has been discovered. Nor is it likely that one will ever be found, for the idea of a one-to-one mapping between genes and behavior is unlikely given what is now known about gene expression, and is implausible in light of the complexity and cultural variation of cooperative behaviors. Thus, when we introduce genetic transmission in our models (as we do in Chapter 10 and Chapter 9), our reasoning operates at the phenotypic level. The ’A allele’ that accounts for altruistic behavior in Chapter 8 is just a phenotypic character that is transmitted exclusively from parent to child, thus abstracting from diploid reproduction, complex gene interactions, the vagaries of development and other aspects of real human genetic transmission, development and phenotypic expression. Similarly, the strategies studied in Chapter 10 are just bi-parentally inherited haploid genotypes. In Chapter 11, where we study the evolution of the human capacity to internalize norms, the ’internalization allele’ is just a behavior acquired from parents.

This phenotype-based approach is a standard tool for the study of the evolution of social behavior in humans and other animals, and has a cogent justification as a device for abstracting from inconsequential complications surrounding the mechanics of genetic inheritance (Grafen 1991, Eshel and Feldman 1984, Hammerstein 1996, Eshel et al. 1998, Levin 2009). Moreover, because it uses observable phenotypes rather than unknown genotypes and developmental processes as the basis for analysis, the approach is readily applied to the kinds of empirical questions we address here.

Second, as is conventional in all models of selection, relative payoffs, whether in terms of fitness, material reward, social standing or some other metric, influence the evolution of the population shares of various behavioral types, higher payoff behaviors tending to increase their frequency in a population. The resulting so-called payoff monotonic dynamic is often clarified using ‘as if’ optimization algorithms, though in doing this we do not attribute conscious optimization to individuals. Nor do we conclude that the resulting outcomes are in any sense optimal. In general they are not. The aggregation of individually optimal choices is universally suboptimal, except under highly unrealistic conditions.

Individuals with higher payoffs may produce more copies of their behaviors in subsequent periods either through the contribution of their greater resources to differential reproductive success or because individuals disproportionately adopt the behaviors of the more successful members of their
group. The latter may occur voluntarily, as when youngsters copy stars, or coercively as when dominant ethnic groups, classes, or nations impose their cultures on subjugated peoples. Of course, cultural transmission may also favor lower-payoff behaviors. We will introduce just such a process when we study socialization and the internalization of norms in Chapter 11.

Would the alternative models common to both biology and economics, kin altruism, reciprocal altruism, and reputation-building, have been sufficient? Even supposing that our Pleistocene ancestors did live in isolated and insular groups of close kin (which we have shown in Chapter 7 to be highly unlikely), these models would fail to explain what is distinctive about human cooperation, namely that it takes place among large numbers of individuals beyond the immediate family.

Third, because positive feedbacks are common in the processes of behavioral and institutional change we study, otherwise identical populations may typically exhibit quite different trajectories, reflecting the multiplicity of equilibria that is typical of models with positive feedbacks. The equilibrium that occurs need not be that with the higher average payoff. The process of selection among equilibria may be on such a long time scale that two populations described by exactly the same model may exhibit dramatically different distributions of behaviors for thousands of generations. The process of determining which of many possible equilibria will occur, termed equilibrium selection, thus assumes major importance.

Fourth, the emergence, proliferation and biological or cultural extinction of collections of individuals, such as foraging bands, ethno-linguistic units, and nations, and the consequent evolutionary success and failure of distinct group-level institutions such as systems of property rights, marital practices, and socialization of the young, is an essential, sometimes the pre-eminent, influence on human evolutionary processes. The maintenance of group boundaries (through hostility towards ‘outsiders’ for example) and lethal conflict among groups are essential aspects of this multi-level selection process. Within-group non-random pairing of individuals for mating, learning and other activities also plays an important part.

Fifth, chance, in the form of mutation, recombination, developmental accidents, behavioral experimentation, deliberate deviance from social rules, perturbation of the structure of social interactions and its payoffs and other stochastic influences, plays an important role. It is often the case that a chance innovation is selectively neutral, not affecting payoffs given the conditions under which it is introduced; in this case it may succeed by chance.
Finally, answers to the question as to how we became such a cooperative species must be empirical. The question is not about “which model works?” They all work, if mathematical coherence is the bar. The question is about something that actually happened in the human past. Thus we measure the empirical plausibility of alternative explanations against the conditions under which early humans lived during the Pleistocene, roughly 1.6 million years before the present until the advent of agriculture beginning about 12,000 years ago, and especially the last 50 or so millennia of this period. Here is Christopher Boehm’s (2007) summary, based on the common characteristics of the 154 foraging societies (about half of those in the ethnographic record) thought to approximate ancestral “highly mobile...storage-free economic systems”:

These highly cooperative nomadic multi-family bands typically contain some unrelated families, and band size, while seasonably variable, seems to be around 20-30 individuals with families often moving from one band to the other. Band social life is politically egalitarian in that there is always a low tolerance by a group’s mature males for one of their number dominating, bossing, or denigrating the others...economic life also tends to be quite egalitarian because of nomadism and a strong sharing ethic which dampens selfish and nepotistic tendencies....regional social networks exist...[and] socially or militarily facilitated group defense of resources is far from infrequent...fueled by ethnocentric tendencies....Drastic resource unpredictability, another likely factor [contributing to group conflict] could have been especially important in the changeable Pleistocene.

We will consider the relevant archeological, climatic, genetic, ethnographic and historical evidence in detail in Chapter 7. Of course, models of the emergence, proliferation and persistence of modern human behaviors must apply to the whole sweep of human history and prehistory as well, including the past 10,000 years.

2.4 Preview

In the next chapter we review a group of models commonly used by biologists to explain cooperation, including kin altruism, reciprocal altruism,
indirect reciprocity, and signaling, showing why these are not adequate to explain the distinctive aspects of human cooperation. In Chapter 4 we ask if recent advances in the theory of repeated games, as exemplified by the so-called Folk Theorem and related models, address the shortcomings of the biological models, finding that they do not. The conclusion is that many of the distinctive aspects of human cooperation, that it takes place in large groups and even among total strangers, cannot be explained by either self-interest with a long time horizon or by the tendency of family members to care for one another.

Fortunately, there is a far simpler explanation of human cooperation: people like to do it or feel that they ought to, and we dislike people who exploit the cooperation of others. In Chapter 5 we review experimental and other evidence showing that even in one-shot interactions many individuals, most in some settings, willingly cooperate with strangers even at a cost to themselves and they enthusiastically punish shirkers who seek to exploit the cooperation of others. Among the social preferences providing the proximate motivation for this cooperative behavior is what we call strong reciprocity, namely a predisposition to help individuals who comply with social norms and to inflict harm on those who do not.

These findings pose the evolutionary puzzle to which the remainder of the book is addressed: how did humans come to have these social preferences? In Chapter 6 we introduce the models of group competition that will play a central role in our explanation and in Chapter 7 we show that prehistoric human society was a social and natural environment in which these so-called processes could have worked with great force (Keller 1999). The next chapter explains how the distinctive institutions of hunter-gatherer society coevolved with a predisposition for altruistic behavior through a process of group competition. Chapter 9 explains the process by which hostility towards outsiders and a warlike disposition could have proliferated and how it strengthened the group competition processes essential to the evolution of altruistic cooperation. Chapter 10 explains how the punishment of individuals who exploit the cooperation of their fellow group members could have initially emerged and proliferated. In Chapter 11 we move from the evolutionary processes accounting for human cooperation to the proximate motives, stressing social emotions such as reciprocity and shame. Chapter 12 turns to deliberate socialization to internalize group norms, why this capacity for internalization could have evolved, and why the norms internal-
ized tend to be group-beneficial. Our final chapter reviews our explanation and considers the future of cooperation.

Human cooperation takes many forms, and these have not evolved by a single mechanism but by many. Because most of the human past approximates the experience of foraging bands more closely than other contemporary or recent societies, we give special attention to cooperation among hunters and gatherers. In this respect, the assessment by Polly Wiessner (in a personal communication) based on her extensive field work among the !Kung in Southern Africa and the Enga in New Guinea captures our understanding of the complexity and diversity of cooperation.

In all the societies where I have worked...kin selection, reciprocal altruism and strong reciprocity...operate in different realms. Kin selection can account for much generalized assistance such as everyday sharing, reciprocal altruism lies at the heart of many long term economic relations that appear altruistic, and strong reciprocity maintains norms that make group living possible. One could have a family group cooperating on the basis of kin selection and many individual partnerships...that are based on reciprocal altruism, but without strong reciprocity to create a common matrix of norms, it seems impossible to have large cooperative communities. Kin selection, reciprocal altruism and strong reciprocity account for different aspects of cooperation, together they make cooperative communities possible.

In the pages that follow we emphasize strong reciprocity and other forms of altruism as proximate motives for cooperation because of the central role that they play in sustaining cooperation.
The Biology of Self-Interested Cooperation

Like successful Chicago gangsters, our genes have survived, in some cases for millions of years, in a highly competitive world. This entitles us to expect certain qualities in our genes. I shall argue that a predominant quality to be expected in a successful gene is ruthless selfishness. This gene for selfishness will usually give rise to selfishness in individual behavior.


Humans are cooperative. But our genes, like Dawkins’ Chicago gangsters, are selfish. This conundrum has challenged biologists to explain the cooperative behaviors of a species that natural selection must have endowed with a predisposition to selfishness in individual behavior. Economists are in practice no less committed to the assumption of universal self-interest, but for different reasons, and have provided complementary models of self-interested cooperation that we will study in the next chapter. In this chapter we review those biological explanations of human cooperation that show that self-interested motives may underlie many cooperative behaviors. In doing this we draw on a remarkable ensemble of papers that revolutionized our understanding of the biology of social behaviors. Dugatkin (2006) is an introduction to the lives and contributions of the leading biologists of social behavior from Kropotkin at the turn of the 20th century to the present.

The first is William Hamilton’s “The Genetical Evolution of Social Behavior” published in 1964, that introduced the concept of inclusive fitness and framed thinking about the evolution of altruism for an entire generation. In 1970, G. R. Price provided a ingenious method for the analysis of selection processes operating at multiple levels. The method was quickly picked up by Hamilton. This was soon followed by Robert Trivers’ 1971 “The Evolution of Reciprocal Altruism,” initiating one of the most vibrant research paradigms in animal behavior to this day, and suggesting some of the distinctive features of human behavior. Two years later Luigi Luca
Cavalli-Sforza and Marcus Feldman published “Cultural versus Biological Inheritance,” initiating the still emerging field of gene-culture coevolution. Almost simultaneously, in 1973, John Maynard Smith produced his most influential paper, coauthored with G. R. Price, entitled “The Theory of Games and the Logic of Animal Conflict.” This paper initiated the field of evolutionary game theory and applied this theory to the dynamics of animal behavior. This cornucopia for the science of social behavior was capped off just sixteen years after Hamilton’s initial contribution by his collaboration with Robert Axelrod, “The Evolution of Cooperation,” (1981) in which Trivers’ insights were formalized using evolutionary game theory and then deployed along with inclusive fitness to provide a convincing explanation of the initial emergence of dyadic cooperative behaviors in a population of erstwhile uncooperative animals.


We will review these models to determine whether the mechanisms they describe explain human cooperation. An empirically relevant model would explain helping behaviors among individuals with moderate, but not unlimited patience, who occasionally make mistakes either in their actions or in their perceptions of what others have done, whose information may thus be inaccurate, and additionally may not be common to all members of a group, and who interact with both close family members and others in groups of substantial size.

Moreover, an adequate model would need to show why a cooperative outcome is both accessible and stable from an evolutionary standpoint, in that it could have emerged in an environment in which the behavior was initially rare, and could have persisted in the presence of environmental and behavioral chance events and challenges from competing behaviors.

By these two standards, empirical relevance and dynamic accessibility and stability, we find that none of the models surveyed singly is entirely adequate, providing at best an implausible account of the evolution of human cooperation. We will see in the next chapter that many models proposed by economists also fail this twin test of empirical relevance and evolutionary stability.
In the next five sections, we review the models. We call the models surveyed here ‘biological’ because they were developed by biologists and have since provided the main fitness-based explanations of the evolution of cooperation. Some of the models, those based on costly signaling and repeated games (§A2) for example, were independently developed by economists and have been used to explain cooperative behavior based on utility rather than fitness maximization. We will not address in this chapter the proximate cognitive and affective processes that would induce an animal to adopt the behavior in question. Though cast in fitness terms the models to follow are of more general application, and are readily extended to treat a process of cultural evolution in which individuals tend to adopt behaviors of others who have higher material payoffs or that would have yielded the individual higher payoffs in the past.

We postpone until Chapter 6 a consideration of models of group competition. There we will also explain the common theme in all of these models, namely that for a helping behavior to proliferate and persist in a population it must either directly confer benefits on the actor sufficient to offset the costs, or social interactions must be structured so that those predisposed to help others must receive more help from others than would occur by chance.

### 3.1 Kin Altruism

Hamilton’s rule for the evolutionary success of altruism (Hamilton 1964) states that in a one-shot (i.e., non-repeated) interaction, conferring a fitness benefit \( b \) on another individual at a direct fitness cost to oneself of \( c \) will be favored by natural selection if

\[
rb > c;
\]

where \( r \) is the genetic relatedness between the actor and the beneficiary (one half for siblings, a quarter for nieces and nephews, an eighth for cousins and so on). When cooperation evolves according to Hamilton’s rule, it increases the expected frequency of the altruistic allele; i.e., the inclusive fitness of the actor is enhanced by an amount \( rb - c \). Rowthorn (2006) offers a clear explanation of how Hamilton’s rule works.

Recalling that helping another is altruistic if the actor would gain higher fitness by declining to help, we note that not all behaviors considered as examples of kin altruism are altruistic. The cost of helping one’s genetic
relatives may be more than compensated by one’s own resulting direct fitness gains, in which case helping is not altruistic (the cost, $c$, in this case would be negative). This is likely to be the case if the relatives in question are one’s children or grandchildren and are thus counted in one’s own fitness. But it is less likely to be the case for other equally close relatives, such as siblings or parents.

Hamilton’s rule captures the key insight common to all models of the evolution of altruism, namely that an altruistic allele cannot proliferate if the bearer is no more likely to receive help from those with whom he or she interacts than would occur by chance. The term kin altruism or kin selection was initially applied to those cases in which the positive assortment indicated by $r$ occurs because individuals interact more frequently than would occur by chance with kin, a term by which we refer to close family members related by recent common descent, more than would occur interactions occurred randomly. Non-kin are pairs of individuals who do not share a recent common ancestor. The terms have since been extended to treat positive assortment arising for any reason. In the multi-level selection models that we address in Chapter 6, for example, assortment occurs because groups are small and reproducibly isolated.

Positive assortment is also at work in the spread of altruism among descendants of a recent common ancestor and its proliferation by multi-level selection. In Chapter 6 and Chapter 7 we will see that group selection models reproduce a variant of Hamilton’s rule. But the two mechanisms differ in important ways; in the forms of behavior likely to be supported, parental care that need not be altruistic at all in the case of kin selection, as opposed to an altruistic willingness to die for one’s group. In the case of kin selection, they differ also in the conditions under which each of these forms of assortment is likely to promote altruism. There is an extended period of dependency and limited dispersal of offspring in the former, as contrasted with frequent inter-group conflict in the latter. Because these differences will emerge as important in our account of the evolution of human cooperation, to avoid confusion we will use the term “kin” to mean close genealogical relatives, so that kin selection is exactly as originally defined by John Maynard Smith in his influential “Group selection and kin selection,” namely, “the evolution of characteristics which favor the survival of close relatives of the affected individual” (Maynard Smith 1964): 1145.

Both kin selection and group selection are based on Hamilton’s concept of inclusive fitness in that both take account both of the direct effects of one’s
actions on the evolutionary success of a trait but also the indirect affects which arise when family structure or group structure result in individuals being more likely than by chance to interact with those bearing the identical trait.

Examples of kin altruism are preferential helping of close kin, which is widely observed in the care of offspring in many animals, and among humans in such diverse aspects of behavior as food-sharing (Gurven et al. 2002, Case et al. 2000), homicide (Daly and Wilson 1988), and migrants’ remittances (Bowles and Posel 2005). However, cooperative behaviors among male chimpanzees, female bonobos, and some other animals are related weakly, if at all, to kin status (Mitani 2006, Gerloff et al. 1999, Silk 2006). About chimpanzees, Vigilant et al. (2001) write: “Kinship may explain cooperative behaviors directed against other communities, but is unlikely to explain the high levels of affiliation and cooperation seen for male within-community interactions.”

Even for intra-household allocations among humans, kin altruism is far from an adequate explanation of helping behaviors. In some studies of food sharing in small-scale societies, kin-altruism effects appear to be very modest or virtually absent (Gurven et al. 2000a, Kaplan and Hill 1985b, Smith et al. 2002). In modern societies important aspects of intergenerational inheritance patterns do not conform to the expectations of the kin altruism fitness model in that the spouse, who is typically genetically unrelated or only distantly so frequently gets a very substantial bequest, and children typically receive equal shares irrespective of their age, health status and other correlates of their reproductive value. A recent study of migrant workers’ remittances of a portion of their urban wages to their rural families of origin in South Africa (Bowles and Posel 2005) shows that less than a third of the remittances sent home can be attributed to the relatedness of the migrant to the members of the household of origin.

But the primary weakness of kin altruism is that it fails to explain the forms of human cooperation that extend considerably beyond the family and even to virtual strangers. The key insight captured by Hamilton’s rule for the study of the distinctively human cooperation is that if altruism is to proliferate among people who are not closely related, the ratio of benefits to costs must be substantial. Thus to explain how quite costly forms of altruism could proliferate when interactions are among people with very modest degrees of genetic relatedness we will need to show that the group benefits of some forms of altruism may indeed be substantial. We will see in sub-
sequent chapters that this indeed was the case where groups competed for survival both against one another and against nature, but rarely otherwise.

3.2 Reciprocal Altruism

Reciprocal altruism, which is long-run enlightened self-interest, the second workhorse model for the explanation of cooperation by biologists, was initially proposed by Robert Trivers (1971) and independently by economist James Friedman (1971). To see how this works, consider a pair of unrelated individuals from a large population who are randomly paired to play the prisoner’s dilemma game repeatedly. In every period the available actions are to help the other at a cost $c > 0$, thus providing a benefit to the other of $b > c$, where $b$ and $c$ are measured in fitness terms, or not to help (defect), incurring no cost and providing no benefit. Clearly, each would gain if each helped, each receiving a net gain of $b - c > 0$. But considering just a single period, as we saw in Chapter 2, defecting is what is termed the dominant strategy: a player earns a higher payoff by not helping, no matter what his partner does.

Suppose the benefits and costs of the interaction are as in Figure 3.1, and the two available strategies for the repeated game are Defect (to not help, with no benefits or costs), or Conditional Cooperation, which involves helping in the first period and in subsequent periods, adopting the strategy played by the other in the previous period (sometimes termed the *nice tit-for-tat* strategy) (the “nice” referring to cooperating in the first period). To study whether helping could evolve in this setting we must first determine if helping individuals (those playing nice tit-for-tat) could persist in a population. In other words, is helping stable? We will presently ask if such an outcome is accessible.

We explore stability by supposing that virtually all members of the population are tit-for-tat helpers and then ask if a mutant defector could pro-
Chapter 3

liferate in this population, eliminating the helpers. If this is not the case, then tit-for-tat is termed an evolutionarily stable strategy and a population of all helpers is termed uninvadable. An evolutionarily stable strategy is a best response to itself, in the sense that no other strategy can do better playing against it. Robert Axelrod and William Hamilton (1981) showed that there are conditions under which Conditional Cooperation will be a best response to itself, so that each of two Conditional Cooperators will do better by helping the other as long as the other does the same, and when paired with a Conditional Cooperator, Defect cannot do better than Conditional Cooperation.

When these conditions are met, we would expect a large population of randomly paired Conditional Cooperators to persist in cooperating. Mutual Defection in this population would also be evolutionarily stable, because Defect is always the best response to Defect. But at least mutual cooperation supported by tit-for-tat players would be stable if ever it were to occur. Assuming that after each round of play the interaction will continue for one more period with probability \( \delta \), the expected duration \( D \) of the game is \( 1/(1 - \delta) \) periods. To see this note that \( D = 1 + \delta D \), where the 1 represent the current play, and \( \delta D \) represents the expected number of future plays; that is, with probability \( \delta \) he interaction continues, and if it does continue to the second round then its expected duration at that point is \( D \), just as it was at start of the first round. In Appendix A3 we show that \( \delta \) can also be interpreted as a time discount factor (that is the reciprocal of one minus what economists term the rate of time preference) reflecting an individual’s impatience or other reasons for a higher valuation on payoffs nearer to the present in an interaction that is repeated forever. There we also show how a probability of continuation less than one can be combined with time discounting.

The result follow by solving this equation for \( D \). To find the condition under which Conditional Cooperation is a best response to itself, we compare the payoff to a Conditional Cooperator when paired with another Conditional cooperator, namely \( b - c \) per period times the expected duration of the interaction \( 1/(1 - \delta) \), with what a Defector would get when paired with a Conditional Cooperator; that is, \( b \) for the first period and then zero thereafter, given that the tit-for-tat playing Conditional Cooperator will respond to the defection by defecting in the next and subsequent periods. Thus Con-
ditional Cooperation will be a best response to itself if
\[
\frac{b - c}{1 - \delta} > b,
\]
(3.2)
or, rearranging,
\[
\delta > \frac{b}{c},
\]
(3.3)
which says that the probability of the interaction continuing must exceed the cost-benefit ratio of the helping action.

There is an equivalent way to derive the expected payoff to playing Conditional Cooperate when paired with a Conditional Cooperator a method that will make use of this derivation when we consider indirect reciprocity. Call this expected payoff \(v\). Then
\[
v = b - c + \delta v.
\]
(3.4)
This expression depends on the fact that if the interaction continues to the second period, then the expected payoff from the second period on is the same as the expected payoff was at the start of the interaction. If we solve (3.4) for \(v\) we get \(v = \frac{b - c}{1 - \delta}\), as expected.

When compared to Hamilton’s rule, condition 3.2 makes it clear that the repetition of the interaction (which happens with probability \(\delta\)) is analogous to genetic relatedness as a support for cooperative behaviors. If this condition holds in a population composed virtually entirely of Conditional Cooperators, and if pairing is random, Defectors will have lower fitness than Conditional Cooperators, so the Conditional Cooperators are not invadable by Defectors.

While mutual defect is also an equilibrium, as Axelrod and Hamilton showed, reciprocal altruism can piggy-back on kin altruism to get started. If the interacting pairs are close family members, for example, and if the termination probability is sufficiently low, Conditional Cooperators can invade an all Defector population, thus accounting for the emergence of cooperation when initially rare. Once Conditional Cooperators becomes are prevalent, their payoff advantage no longer requires relatedness among the interacting pairs.

Reciprocal altruism paired with kin altruism thus provides a simple and elegant example of a mechanism that could account for the emergence and persistence cooperation among empirically plausible human actors.
course, the above argument is hardly definitive because we have not analyzed what happens if there are behavioral or perception errors, and we have not considered invasion by strategies other than Defect. It turns out that in dyadic interactions, the Conditional Cooperator strategy is quite robust to such adverse conditions, because it has three attractive characteristics: it is nice in the sense that it never defects first, it is punishing in that it does not let an opponent get away with not helping without paying a penalty, and it is forgiving, in that if a partner who has defected returns to cooperation, the Conditional Cooperator will return to cooperation as well. However, other strategies with these three characteristics can often out-compete the Conditionally Cooperators (Nowak and Sigmund 1993).

Many biologists have considered reciprocal altruism to be a common form of cooperation among non-kin in other animals. Because cooperation based on reciprocal altruism was thought to be common in the non-human animal world, it was natural for biologists and social scientists to accept it as the standard explanation for human cooperation as well. However, there is little evidence of reciprocal altruism in non-human animals. Peter Hammerstein (2003) writes, “After three decades of worldwide research on reciprocal altruism and related phenomena, no more than a modest number of animal examples have been identified.” A major impediment is that reciprocal altruism requires that animals act as if future rewards are almost as highly valued as those in the present. This is virtually never the case (Noe 1990, Clements and Stephens 1995, Connor 1995, Hammerstein 2003, Whitlock et al. 2007, Russell and Wright 2008, Clutton-Brock 2007, Connor 2010). Behaviors that at first appeared to be reciprocal have frequently, on further study, been better explained as simple mutualism in which the benefit to the actor compensates for the cost of the action irrespective of the action taken by the other (Dugatkin and Mesterton-Gibbons 1996, Milinski 1996, Stephens et al. 2002).

Non-human primates may eventually provide exceptions to this generalization. There is some evidence (de Waal and Davis 2003) that reciprocity plays a role in alliance-formation among macaques and chimpanzees, in the latter among unrelated individuals in high-risk power struggles, both in the wild as well as among captive animals. Capuchin monkeys have been observed experimentally to share more food with a partner whose assistance was needed to acquire the food (de Waal 2000) and chimpanzees to give food to those who groomed them in the past (de Waal 1997).
Among humans the experimental and field evidence for the contribution of repeated interactions to cooperation is overwhelming and well known. An experiment by Gächter et al. (2004) suggests its importance. Four hundred and fifty six Swiss students were randomly assigned to two roles similar to those of employer and employee in a natural setting. The ‘employer’ offered a wage, a deduction from his profits, and the ‘employee’ responded with a level of ‘effort,’ which was costly to provide. The contrast between two treatments that the experimenters implemented is instructive. In the stranger treatment the pairs were shuffled every period, so that each period was a one-shot interaction for the participants, who were certain they would not encounter any partner more than once. In the partner treatment the two remained paired over ten periods, and this set of ten periods was itself repeated three times. The dominant strategy for a self-regarding ‘employee’ in the stranger treatment (which, recall, meaning the response giving the highest payoff no matter what the other player does) was to offer one unit of effort. In fact, ‘employers’ made wage offers far more generous than the minimum required to illicit the one unit of effort (we explore this phenomenon at greater length in §5.3). Figure 3.2 gives the effort responses by ‘employees in the two treatments over time.

Figure 3.2. Repetition Supports Cooperation in a Gift Exchange (Wage-Effort) Experiment. Based on Gächter, Kessler, and Königstein (2004)
Strikingly, the effort offered in the stranger treatment is much higher (four times higher) than would have been optimal for a self-regarding ‘employee’. But more important for the questions under consideration here, the repeated interaction resulted in much higher levels of effort than the stranger treatment, and effort rose not only over the three sets of treatments but also, except for a dramatic endgame drop-off, within the three sets of play. The fact that repetition contributed to cooperation and that the subjects readily understood the difference between repeated and non-repeated interactions is evident by the sharp reduction during the last two periods of play. The endgame drop-off cannot be due to learning, as high levels of effort are restored and even exceeded when the second and third set are initiated. The fact that in the last period effort did not fall to the level of the stranger treatment could have occurred because, while repetition engaged the self-regarding motives stressed by the reciprocal altruism model, it also tapped social preferences that the stranger treatment did not evoke.

Among humans in dyadic interactions, reciprocal altruism provides a convincing explanation of at least some of the cooperation we observe (Cooper et al. 1996, Keser and van Winden 2000). Many forms of cooperation in our ancestral past, judging by the practices of foragers in the ethnographic record, had precisely this structure. The sharing of some food in foraging groups does not take the form of the redistribution of food that has been pooled on a group wide basis, what we term common pot redistribution, but rather is network based (Kaplan and Gurven 2005). Thus, sharing may take the form of reciprocation in dyadic relationships. In these cases, repeated interactions that reward cooperators with higher fitness or other payoffs make helping a form of mutualism, as in the model of Axelrod and Hamilton (1981).

### 3.3 Reciprocal Altruism in Large Groups

The plausibility of the model, however, does not extend to large groups in which members sometimes make mistakes. Interactions of this kind are common and certainly were so our ancestral past, including such fitness relevant activities as common defense and predation, sharing information, cooperative hunting and gathering, and sustaining social norms through the punishment of transgressors.

Because reciprocal altruism models are dyadic, it is not even clearcut how one would extend them to groups with more than two players. In this sec-
tion we consider the most plausible extension: individuals cooperate in a repeated public goods game with $n > 2$ players if a sufficient number of other players cooperated on the previous round. In the one-shot $n$-player public goods game, each player has two strategies: Contribute and Defect. Contributing generates a benefit $b$ shared equally by all other members of the group, at a cost $c < b$ to the Contributor. All players decide simultaneously whether to contribute or defect, after which the game is over. Clearly, if all members contribute, each earns $b - c > 0$, but each individual has an incentive to defect, in which case his payoff is greater by the amount $c$ than his payoff to contributing, no matter what the other players do. Hence, in a one-shot public goods game, self-regarding agents will all defect.

By repeating the public goods game many times, with a fixed probability of ending after each round, however, cooperation can be sustained under favorable conditions. But these conditions are quite stringent. We will show that this game supports cooperative outcomes only if the group is small, the returns to cooperation are high, the behavior of each group member is known with a high degree of accuracy by the other group members, errors in execution are infrequent, and group members are very patient and interactions typically endure for many periods.

The reason for the ineffectiveness of reciprocal altruism for groups with several members is simple. In groups of two, a free rider cannot go undetected and when one member defects in order to punish a Defector, the punishment is uniquely targeted on the Defector. But, in groups larger than two, a retaliatory defection punishes not only the initial defector, but also all other members of the group. Moreover, other group members may not have observed the initial defection and hence may think that a retaliatory defection is a free-riding defection, inviting further retaliatory defections, as we shall see occurs in an experimental public goods game with human subjects in Chapter 5 (Figure 5.2).

To illustrate the consequences of extending the reciprocal altruism model to groups larger than two, we will develop an agent-based model. We explore the shortcomings of analytical models of cooperation in the public goods game in the next chapter, including cases like the public goods with punishment game, which we explore experimentally in Chapter 5. In this game, punishment can be targeted on particular miscreants, thus rendering punishment an effective disciplining device. We explain the structure and logic of agent-based models in Appendix A1.
Suppose a large population consists of $N$ groups of $n$ members each, and each group plays a public goods game repeatedly $d$ times. We will call this series of $d$ rounds an encounter. At the end of each encounter, players reassort randomly into new groups of size $n$ and carry out another encounter. By cooperating, a player confers a benefit of $b$ on the other members (i.e., a benefit of $b/(n - 1)$ per other member) and a cost $c$ to himself. We generalize the Conditional Cooperator strategy to groups of size $n > 2$ by assuming that there are $n + 1$ possible types of players, called $t$-Cooperators, for $t = 0, \ldots, n$. A $t$-Cooperator cooperates in the current round provided at least $t$ other players cooperated in the previous round. We call an $n$-Cooperator, who never cooperates, a Defector, and we call a 0-Cooperator an unconditional Cooperator. On the first round, all players apply the $t$-criterion as though everyone cooperated in the previous round (i.e., all types cooperate except the $n$-Cooperator, who defects on all rounds). Finally we assume that a player who attempts to cooperate will accidentally defect with probability $\epsilon > 0$. We call this an execution error, and we call $\epsilon$ the execution error rate. Note that when $n = 2$, our model reduces to standard conditional cooperation, where the Conditional Cooperators is a 1-Cooperator, universal defect is a 2-Cooperator, and an unconditional cooperator is a 0-Cooperator.

Our central question will be the frequency of cooperation that can be sustained in the long run of this system, for different choices of the benefit $b$, error rate $\epsilon$, duration of the encounter $d$, and the group size $n$. We created a population with 25 groups of size $n = 2, 4, 6, 8, 10, 12, 14$ playing a public goods game repeatedly for twenty five periods ($d = 25$), in which by cooperating, an individual contributes $b$ to the other players at a cost of $c = 1$ to himself, where $b = 2$ and $b = 4$. We begin (initialize) the simulation by assigning to each individual a value of $t$ in the range $t = 0, \ldots, n - 1$ with equal probability. The exact initial conditions are unimportant because a small mutation rate ensures that no strategy becomes completely extinct, and in fact the simulation attains its long-run dynamic, called the model’s stationary distribution, in the very short time of several hundred periods, as long as some of each type is present from the start.

After each encounter, 5% of individuals are replaced by new individuals, using a Darwinian fitness criterion according to which the probability of reproduction is proportional to the individual’s payoff relative to others in the entire population. We assume a mutation rate of 2% per newly-created individual, which means there is one mutation about every 50 encounters.
Also, we assume an execution error rate varying from $\epsilon = 0\%$ to $\epsilon = 10.0\%$. To promote cooperation in the face of errors, we assume that an individual who accidentally defected (we assume this fact is known to all group members, so it is public information) cooperates unconditionally on the next two rounds, thus allowing cooperation to be restored (this behavior is known as contrite).

As seen in Figure 3.3, if $n = 2$ and $b/c = 2$, if the error rate is zero, full cooperation results. However, intuitions based on the working of this dyadic case and zero error rates do not extend to larger groups and plausible error rates. Even at a group size of $n = 4$ and an error rate of only 2%, very little cooperation occurs. When $b/c = 4$, we see in the right pane of Figure 3.3 that cooperation can be sustained for $n = 4$, but for larger group sizes, and at error rates of 4% or more, very little cooperation results. This shows clearly that reciprocal altruism need not support cooperation in groups of size larger than two unless errors are virtually absent.

These simulations assume a very high probability that the interaction will be continued: $\delta$ from equation 3.2 is 0.96, close to the maximum of what is plausible if a ‘period’ is a year, given human mortality rates, even setting aside other reasons for devaluing the future and terminating interactions, such as myopia and migration. Simulations with more empirically plausible assumptions concerning $\delta$ give even more negative results.

But there is even more bad news when it comes to applying this model to real populations. We have assumed that all group members receive the same information indicating the cooperation or defection of each other group member. That is, we assume public signals. Moreover, we have assumed that the only kind of error is executional in the sense that an individual tries, but fails to cooperate. It is instructive to move from public to private signaling, in which each group member receives an independent signal with error rate $\epsilon$ as to whether each other member contributed. We call such errors perceptual. In this new setting, an individual can have an execution failure rate of $\epsilon$, and in addition, a cooperating individual may be perceived as having defected also with probability $\epsilon$, independently, by each other group member (we use the same error rate in both cases for ease of exposition). Our simulations of this situation, illustrated in Figure 3.4, show that virtually no cooperation is possible for $n > 2$ and $\epsilon \geq 2\%$ when $b/c = 2$, and very little is possible even for $n > 4$ or $\epsilon > 6\%$ when $b/c = 4$. 
Figure 3.3. The Rate of Cooperation in the Reciprocal Altruism Game with Public Signals for various Group Size and Error Rates, Public Signals. The Net Cooperation Rate is the mean fraction of individuals who attempted to cooperate on a single round. Each encounter lasted 25 game repetitions, so the implicit discount rate is 4%. Each bar represents the average over the final 10,000 periods of 20 independent simulations of 25,000 encounters. Each simulation used 20
Figure 3.4. Cooperation in the Reciprocal Altruism Game with Private Signals for various Group Size and Error Rates. The parameters are the same as in the previous figure.
Students of economic theory may be surprised by these results, because the Folk Theorem of repeated game theory appears to assure us that for any group size and any error rate, cooperation may be sustained if the discount factor is sufficiently high. In the next chapter we show that this is not the case when we allow information to be private or we require that cooperation be sustained in a stable equilibrium.

3.4 Reputation: Indirect Reciprocity

In a sizeable group, an individual may interact frequently with a large number of partners, but infrequently with any single one. If the time lag between encounters with a given individual is sufficiently long, unless individuals are infinitely patient and have good memories, the conditions for reciprocal altruism may not obtain.

An alternative mechanism of cooperation is for each individual to keep a mental account of group members who cooperated with their partners in the previous period and those who did not. Let us call an individual who cooperated in the previous period in good standing, and specify that the only way an individual can fall into bad standing is by defecting on a partner who is in good standing. Thus, an individual can always defect when his partner is in bad standing without losing his good standing status. In this more general setting the Conditional Cooperator strategy is replaced by the following standing strategy: cooperate if and only if your current partner is in good standing, except that if you accidentally defected the previous period, cooperate this period unconditionally, thereby restoring your status as a member in good standing.

This standing model, due to the economist Robert Sugden (1986), is an ingenious formalization of the biologist Richard Alexander’s (1987) indirect reciprocity notion. As in the case of conditional cooperation, we need to discover the conditions under which the standing strategy is a best response to itself, so that in a population following the standing strategy, no individual has an incentive to defect intentionally. As before, individuals are paired randomly to play with the same payoffs as before. We assume that after each play the game continues with probability \( \delta \), with \( 0 < \delta < 1 \), and there is a probability \( \epsilon \) of accidentally defecting. Suppose all individuals follow the standing strategy, and now consider an individual in good standing. He receives \( b \) with probability \( 1 - \epsilon \), because with probability \( \epsilon \) his partner fails to deliver the benefit, he pays \( c \). Then with probability \( \delta \),
the game is repeated, in which case with probability \( \epsilon \) he is in bad standing, because he accidentally defected the previous period, and with probability \( 1 - \epsilon \) he is in good standing. Thus, in a population of individuals who have adopted the standing strategy, the expected value of being in good standing is given by

\[
v_g = b(1 - \epsilon) - c + \delta(\epsilon v_b + (1 - \epsilon)v_g),
\]

where \( v_b \) is the expected value of being in bad standing. Analogously to equation 3.4, this equation says that in the first period the payoff is \( b(1 - \epsilon) - c \), and if the interaction continues, which occurs with probability \( \delta \), the expected payoff from then on is the weighted average of the payoffs to being in bad and good standing, the weights respectively being the probability that the player has mistakenly defected (\( \epsilon \)) and its complement.

If an individual is in bad standing, he receives \( b \) with probability \( \epsilon(1 - \epsilon) \), because with probability \( \epsilon \) his partner is in bad standing and delivers the benefit to restore good standing with probability \( 1 - \epsilon \), he pays \( c \) to restore his good standing, and with probability \( \delta \), the game is repeated, in which case with probability \( \epsilon \) his is in bad standing, because even though he paid \( c \), he mistakenly defected, and with probability \( 1 - \epsilon \) he is in good standing. Thus, in the same population, the expected value of being in bad standing is given by

\[
v_b = b\epsilon(1 - \epsilon) - c + \delta(\epsilon v_b + (1 - \epsilon)v_g).
\]

If we solve the previous two equations simultaneously for \( v_g \) and \( v_b \), we find

\[
v_g = \frac{b(1 - \epsilon(1 + \delta)) - c + b\delta \epsilon^2(2 - \epsilon)}{1 - \delta}, \quad (3.5)
\]

\[
v_b = v_g - b(1 - \epsilon)^2. \quad (3.6)
\]

The numerator in this expression for \( v_g \) is approximately (with an error on the order of \( \epsilon^2 \)) the net expected payoff per period, and \( 1/(1 - \delta) \) is the expected number of periods in the game. The loss due to the need to punish execution errors is approximately \( b\delta \epsilon \) per individual per period, which is not large unless the error rate is quite high. It is a best response to cooperate as long as \( v_g \geq b(1 - \epsilon) + \delta v_b \), where the right hand side is the present value of defecting now and hence of falling into bad standing. This inequality simplifies to

\[
\frac{b\delta(1 - \epsilon)(1 + \delta(1 - \epsilon)^2 + \epsilon^2)}{1 + \delta} \geq c. \quad (3.7)
\]
For any given discount factor \( \delta < 1 \), this inequality is satisfied for sufficiently small error \( \epsilon > 0 \), as long as \( b > c \). Ignoring terms in \( \epsilon^2 \) or higher because for small error rates they will be very small, the above inequality simplifies to \( b\delta(1 - \epsilon) > c \). Note the similarity to Hamilton’s rule (3.1), with the degree of relatedness \( r \) replaced by the (generally much larger) \( \delta(1 - \epsilon) \).

Like the reciprocal altruism model, this result is restricted to dyadic interactions and hence cannot explain many important kinds of cooperation. But Parthik Panchanathan and Richard Boyd (2004) have proposed an ingenious use of indirect reciprocity as an adjunct to the \( n \)-player public goods game. Suppose a group of size \( n \) plays an indirect reciprocity game using the standing strategy with execution error rate \( \epsilon \geq 0 \) as described above, and also plays a public goods game. They start with the one-shot public goods game, at the end of which they play the indirect reciprocity game, except individuals start out in good standing if and only if they cooperated in the public goods game. At the end of the indirect reciprocity game, they repeat the sequence of a one-shot public goods game followed by an indirect reciprocity game, indefinitely.

To see why this works, suppose the benefit to others and cost to himself in the public goods game are \( b_g \) and \( c_g < b_g \), while in the indirect reciprocity game, the cost of helping is \( c \) and the benefit to the recipient is \( b > c \). Panchanathan and Boyd (2004) show that, so that as long as

\[
\frac{b(1 - \epsilon) - c}{1 - \delta} \geq c_g.
\]

contributing in the public goods game and adopting the standing strategy in the indirect reciprocity game is a best response to itself, so cooperation can be sustained. Parameters favoring this solution are that the cost \( c_g \) of cooperating in the public goods game be low, the probability of repeating the indirect reciprocity game \( \delta \) is near unity, and the net benefit \( b(1 - \epsilon) - c \) of cooperating in the indirect reciprocity game be large.

The public goods game has two serious weaknesses that the Panchanathan-Boyd model repairs. First, in some cases it is played very infrequently, such as in case of war with other groups, response to famine, and the like. In this case, the probability of repetition in a single time period would be very low, so the cost of defecting would be low as well. Second, as we saw earlier, the only way to punish Defectors in the public goods game is for Cooperators to withdraw their contributions, thus
hurting all other players, not only the Defector. The Panchanathan and Boyd solution is for group members to refuse to cooperate in everyday social exchange with Defectors in the public goods game. In this way, punishment of defection becomes targeted, frequent, and costless to the punisher (he avoids the cost of extending help). Everyday social exchange is thus represented as an indirect reciprocity game, with (3.8) being the condition for the game to foster cooperation in the public goods game.

The major weakness of this model, which it shares with other indirect reciprocity models, is that the standing strategy has very demanding informational requirements. Its ability to target defectors makes it an improvement on the repeated game models of the previous section, but this comes at a steep price. Each individual must know the current standing of each member of the group, the identity of each member’s current partner, and whether each individual cooperated or defected against his current partner, since this information is necessary to ascertain the status of one’s partner in the indirect reciprocity game. But real-world dyadic interactions are often private, and hence are unlikely to be directly observed by more than a small number of others, vitiating the model for groups of any significant size. Because the repeated game models of the previous section punishment is not directed at defectors, they require only that an individual know how many of his current partners defected in the previous period.

Especially problematic in the case of indirect reciprocity is the fact that if the dyadic interactions involved are at all complex, it will be virtually impossible for an outsider to ascertain whether an individual’s action is or is not a defection. Hence, in many cases, the standing of one’s current partner will be subject to considerable error. Equally important, if individuals are entirely self-regarding, they have no reason to report truthfully what they have observed. Though an active area of research, there exist no explanations of how private information could be converted to accurate public information in a population of self-regarding individuals. We return to this question at the close of the next chapter.

We conclude that indirect reciprocity may sustain cooperation in small groups with especially high quality information concerning the private interactions of pairs of individuals. Given the implausible informational demands of these models, it may be surprising that behavior consistent with the indirect reciprocity mechanisms identified by Alexander are common in human society.(Engelmann and Fischbacher 2009, Rockenbach and Milinski 2006). But, this is likely due to the importance of character virtues
(e.g., honesty) and other-regarding preferences, and not entirely to the self-regarding preferences assumed in the indirect reciprocity model.

### 3.5 Altruism as a Signal of Quality

Another form of mutualistic cooperation is that which results when a cooperative act is a difficult-to-fake signal of some characteristic of the actor, such as strength, skill or bravery, that is otherwise difficult to observe. In this case a cooperative behavior may be favored in selection because it enhances the individual’s opportunities for mating and coalition building. This would be the case, for example, if sharing valuable information or incurring dangers in defense of the group were taken by others as an honest signal of the individual’s otherwise unobservable traits as a mate or political ally. In this case self-regarding individuals might engage in group-beneficial activities in anticipation of reproductive, political, or other benefits. Cooperative behaviors would thus result in advantageous alliances for those signaling in this manner, and the resulting enhanced fitness or material success would then account for the proliferation of the cooperative behaviors constituting the signal. Models of this process were developed by the economist Michael Spence (1973), initially applied to educational attainment as a signal, and by the biologist Amos Zahavi (1975) to explain helping behavior in Arabian babblers.

Costly signaling has been proposed as an explanation for some types of food-sharing in human societies, such as providing game that is large and/or difficult to harvest, or large quantities of food for consumption at ritual feasts (Boone 1998, Gurven et al. 2000b, Hawkes et al. 2001, Smith and Bliege Bird 2000, Sosis 2000).

With Eric Alden Smith, we modeled this process as a multi-player public goods game that involves no repeated or assortative interactions, so that non-cooperation would be the dominant strategy if there were no signaling benefits (Gintis et al. 2001). We showed that honest signaling of underlying quality by providing a public good to the rest of the group can be evolutionarily stable, provided that certain plausible conditions hold. The absence of such costly signaling by cooperative behavior is also an equilibrium.

A signaling equilibrium, however, does not require that the signal confer benefits on other group members. Anti-social behaviors could perform the same function, beating up one’s neighbor can demonstrate physical prowess just as much as bravely defending one’s group. If signaling is to be an expla-
nation of group beneficial behavior, we must explain why group-beneficial signaling is favored over antisocial signaling. There are several possible reasons. It may be that the level of public benefit provided is positively correlated with the individual benefit the signaler provides to those who respond to the signal. For instance, the signaler who defends the group may be more likely to confer a benefit (say, protection) on his partner or allies than the signaler who beats up his neighbor. Group-beneficial signals such as sharing one’s prey may attract larger audiences than anti-social signals. Finally, competition among groups for material or fitness resources would favor groups at group-beneficial signaling equilibria over those either at non-signaling equilibria or those at anti-social signaling equilibria. Thus, the effects of signaling and group competition on cooperation may be synergistic rather than simply additive. Group competition provides a reason why the signalling that we observe tends to be group beneficial, while signaling theory provides a reason why signaling of any kind may be evolutionarily stable in a within-group dynamic, thus contributing to between-group differences in behavior and, as we will see presently, thereby enhancing the force of group competition (Bergstrom and Lachmann 2001).

3.6 The Limits of Mutualistic Cooperation

In this chapter we have reviewed four biological explanations of cooperation: kin altruism, reciprocal altruism, indirect reciprocity, and costly signaling. While important in explaining many aspects of human cooperation, these theories do not provide an adequate account of the phenomenon. Kin altruism is important in relationships among close genetic relatives. But other than possibly being the template for more generalized forms of altruism towards non-kin, it does not explain one of the most distinctive aspects of human cooperation, namely that taking place beyond the family.

Reciprocal altruism, which is of course not really altruism at all, provides a plausible account of cooperation among dyads, but not in large group interactions or in the presence of realistic levels of errors of execution or perception, and not when future interactions are unlikely.

Indirect reciprocity facilitates cooperative behavior when interactions are dyadic but it does not support cooperative outcomes except under quite implausible assumptions concerning what individuals can know and remember about the behavior of members of their group when interacting with others (see A4).
Chapter 3

The costly signaling model explains why helping behaviors might not be selected against within groups but it does not provide very convincing reasons to think that helpful signals would be more common than antisocial ones.

Three of these models, reciprocal altruism, indirect reciprocity, and costly signaling, identify the conditions under which the fitness cost of helping others is more than offset by the corollary fitness benefits for the actor. The evolution of helping behavior is possible only when these conditions obtain, and when this is the case, by these mechanisms helping is not altruistic, because the actor would suffer reduced fitness by not helping. These models explain how helping behaviors that seem altruistic when considered narrowly, ignoring the effects of helping behaviors on future interactions, may evolve because they are not altruistic when a more comprehensive view of direct benefits to the actor is taken. Similar reasoning applies to some (but not all) cases of kin altruism.

Our analysis of the shortcomings of standard biological approaches in the preceding pages is based on the simplest of models. Over the past two decades economists have made extraordinary advances in repeated game theory, allowing a new generation of models.

To these we now turn.
Cooperative Homo economicus

It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest.

Adam Smith, *Wealth of Nations* (1937[1776]), Book 1, Chapter 2

Two neighbors may agree to drain a meadow, which they possess in common; because 'tis easy for them to know each other's mind; and each must perceive, that the immediate consequence of his failing in his part, is the abandoning of the whole project. But 'tis very difficult and indeed impossible, that a thousand persons shou'd agree in any such action; it being difficult for them to concert so complicated a design, and still more difficult for them to execute it.


In contrast to biology, where cooperative behaviors have become a central research focus only in recent decades, a major goal of economic theory since its inception two and a half centuries ago has been to explain the mutual benefits provided by a widespread form of voluntary cooperation, namely, mutually beneficial market exchange among self-regarding individuals. This endeavor culminated in the *Fundamental Theorem of Welfare Economics* (Arrow and Debreu 1954, Debreu 1959, Arrow and Hahn 1971), sustaining Smith’s insight that self-regarding behaviors might support socially valued economic outcomes. In the resulting model of exchange, individuals maximize their utility given a set of market-determined prices over which they exercise no control. They thus interact with a list of prices, not with one another. As a result the essential condition for strategic behavior, recognition that the payoffs of each depend on the actions of others is absent. The critical assumption implying that individuals are content to interact through prices alone, and hence have no incentive to engage in per-
sonal interactions, is that all relevant aspects of exchanges can be covered by complete contracts, enforceable at no cost to the exchanging parties.

A second major thrust of economic theory eschewed these implausible “complete contracts” assumptions, and developed models in which the outcomes of exchanges are determined by punishment, threats, and other forms of strategic interaction. These game-theoretic models refine and extend the insights of Shubik (1959), Trivers (1971), Taylor (1976), and Axelrod and Hamilton (1981) that retaliation against defectors by withdrawal of cooperation may enforce cooperation among self-regarding individuals. This literature culminates in the folk theorems of Fudenberg, Levine, Maskin, and others (Fudenberg and Maskin 1986, Fudenberg et al. 1994). A virtue of these models, in contrast to the older non-strategic paradigm in economics, is that in recognizing the ubiquity of incomplete or unenforceable contracts, they describe the real world of interactions among most animals, including humans (Blau 1964, Gintis 1976, Stiglitz 1987, Tirole 1988, Laffont 2000, Bowles and Hammerstein 2003).

But do these models provide an adequate explanation of cooperation among self-regarding individuals? In this chapter, we will show that while the insight that repeated interactions provide opportunities for cooperative individuals to discipline defectors is correct, none of these models is successful. The reason is that even presupposing extraordinary cognitive capacities and levels of patience among the cooperating individuals, there is no reason to believe that a group of more than two individuals would ever discover the cooperative Nash equilibria that the models have identified, and almost certainly, if it were to hit on one, its members would abandon it in short order (a Nash equilibrium is a strategy for each player such that no player could gain by unilaterally changing his strategy). Except under implausible conditions, the cooperative outcomes identified by these models are neither accessible nor persistent. We term them dynamically irrelevant Nash equilibria.

4.1 Folk Theorems and Evolutionary Dynamics

All folk theorems are based on a stage game, that is, an interaction like the one in Figure 3.1, played an indefinite number of times, with a constant, strictly positive, probability that in each period the game will continue for an additional period (§A2). The restrictions on the stage game tend to be minimal and rather technical, so that the public goods game and other social
dilemmas involving costly cooperation fall within the aegis of the major folk theorems (Fudenberg and Maskin 1986, Fudenberg et al. 1994). Players receive signals concerning the cooperate vs. defect behavior of other players in previous periods, and reward or punish other players in such a manner as to render cooperation a self-regarding best response for all players.

The signals concerning the actions of other players may be perfect (completely accurate) or imperfect (inaccurate with positive probability). The signals can also be either public (the same signal received by all individuals) or private (different individuals get different signals, or no signal at all)—see Figure 4.1. As in the previous chapter, we may think of imperfect public signals as caused by perceptual errors, which are then seen by all other players, while private signals are caused by limited scope, in which players observe the behaviors of only a subset of their group members, or perceptual errors, in which a player who has cooperated is signaled as having cooperated to some subset of players, and signaled as having defected to the remaining players. Note that with private signals, a player who cooperates in a given period does not know which players, if any, received an inaccurate signal indicating that he defected.

<table>
<thead>
<tr>
<th></th>
<th>Perfect</th>
<th>Imperfect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public</td>
<td>Accurate signals received by all</td>
<td>Identical noisy signals received by all</td>
</tr>
<tr>
<td>Private</td>
<td>Accurate signals received by some</td>
<td>Different noisy signals received by different people</td>
</tr>
</tbody>
</table>

Figure 4.1. Information structures. Most models of cooperation assume perfect public information, but in all but the simplest human interactions, some critical information is both private and imperfect.

With either perfect or public imperfect signals, a folk theorem can be proved, asserting that any feasible allocation of payoffs to the players that exceeds some minimal amount can be achieved, or approximated as closely as desired, as the equilibrium per-period payoff to the repeated game, for some discount factor strictly less than unity (Fudenberg and Maskin 1986, Fudenberg et al. 1994), and for sufficiently small signal errors.
For example, suppose Alice and Bob play the prisoner’s dilemma shown in Figure 4.2. Of course, in the one-shot game there is only one equilibrium in which both parties defect. However suppose, as in our discussion of reciprocal altruism in the previous chapter, Alice and Bob play this stage game at times $t = 0, 1, 2, \ldots$ This is then a repeated game, in which the payoff to each is the sum of the payoffs over all periods, where following each period the interaction continues with probability $\delta$, with $0 < \delta < 1$.

A strategy in this game that dictates following one course of action until a certain condition is met, and then following a different strategy for the rest of the game is called a trigger strategy.

We can show that the cooperative solution $(5,5)$, which is unattainable in the stage game with self-regarding players, can be attained in the repeated game if $\delta$ is sufficiently close to unity, and each player uses the trigger strategy of cooperating as long as the other player cooperates, and defecting forever if the other player defects on one round. We use the fact (see § 3.2) that for any discount factor $\delta$ with $0 < \delta < 1$, the expected duration of the game in number of rounds is $1/(1-\delta)$.

Now suppose both players use the trigger strategy. Then, the payoff to each is $5/(1-\delta)$. Suppose a player tried to do better by using a different strategy. This must involve defecting forever; for once the player defects, his opponent will defect forever, the best response to which is to defect forever. A player who defects receives 8 immediately and zero in each succeeding period. Thus the defect strategy is superior to the cooperate strategy if and only if $5/(1-\delta) < 8$, or $\delta < 3/8$. Thus, when $\delta > 3/8$, the pair of trigger strategies cannot be invaded by defection.

In fact, there are lots of other equilibria to this game. For instance, consider the following trigger strategy for Alice: alternate $C, D, C, \ldots$ as long

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$C$</td>
</tr>
<tr>
<td>Alice</td>
<td>5,5</td>
</tr>
<tr>
<td></td>
<td>8,-3</td>
</tr>
</tbody>
</table>
as Bob alternates $D, C, D, \ldots$. If Bob deviates from this pattern, defect forever. Suppose Bob plays the complementary strategy: alternate $D, C, D, \ldots$ as long as Alice alternates $C, D, C, \ldots$, and if Alice deviates from this pattern, defect forever. Neither player can increase his payoffs by deviating from his assigned strategy for $\delta$ sufficiently close to unity. To see this, note that if it is profitable to deviate, it will be most profitable for a player to defect when called upon to cooperate, in which case the defector has payoff zero in the first and all succeeding periods, since deviating by cooperating when the strategy calls for defection results in a payoff of $-3$ on the first round and zero thereafter. The payoff to sticking to the trigger strategy is $-3, 8, -3, 8, \ldots$. To evaluate the expected payoff of this infinite stream of returns, note that the player gets $-3$ in the first period, $8$ in the next period, and then with probability $\delta^2$ gets to play the game all over again starting two periods from today. Thus, if $x$ is the expected payoff of the game, we have $x = -3 + 8\delta + \delta^2 x$. Solving this, we get $x = (8\delta - 3)/(1 - \delta^2)$. The trigger strategy beats defection when $x > 0$, which requires that $8\delta - 3 > 0$, or $\delta > 3/8$. Once again, if $\delta > 3/8$, the trigger strategy cannot be beat.

Figure 4.3. **The Folk Theorem.** Any point in the region $OABC$ can be sustained as the average per-period payoff in an equilibrium of the repeated game based on the Prisoner’s Dilemma stage game depicted in Figure 4.1.

The possible expected per-period payoffs in the repeated game based on this stage game are exhibited in Figure 4.3. The folk theorem says in this case that the use of trigger strategies of the kind just illustrated can sustain
any point in the region $OABC$ as an equilibrium of the stage game, provided the discount factors of the players are sufficiently close to unity. For instance, it is easy to show that to attain point A, Alice can use a trigger strategy in which she plays Cooperate in every period but Bob is permitted to play Defect in seven periods out of every seventeen, playing Cooperate in the other ten. If Bob cooperates less frequently than this, Alice pulls the trigger and defects for as long as the game lasts. Bob’s trigger strategy is the mirror image of Alice’s. If the discount rate is $\delta = 0.98$, Alice earns zero on average and Bob earns a bit less than seven.

From this example one suspects that the result may be a little too good to be true, for how would Alice and Bob come to coordinate on this particular pair of trigger strategies? There is an infinite number of other pairs of strategies that would implement the other points along the frontier $ABCD$. Why, for example, would Alice not seek to turn the tables by preemptively adopting Bob’s trigger strategy, the best response to which for Bob would be to adopt its complement, implementing point $C$ in the figure, resulting in a reversal of the payoffs at point $A$.

When we ask how players might actually come to coordinate on an efficient pair of trigger strategies (that is, one on the frontier $ABCD$) or for that matter, any set of trigger strategies at all other than mutual defect, the theorem is silent. But when groups are large, the problem is far worse than is suggested by the difficulty that a single pair of individuals, Alice and Bob, might have in coordinating on a set of strategies, for as Hume observes about just this kind of problem in the headquote of this chapter; it is a lot easier to come to a working arrangement among two individuals than among a thousand independent spirits, in the absence of a regulating social institution. To understand just how serious this coordination problem is, we need to consider what happens when the numbers involved are considerable, and the information each has on the others actions is subject to error.

### 4.2 The Folk Theorem with Imperfect Public Information

The success of the model in the previous section depends critically on the fact that the signal is public and perfect. Errors can have disastrous consequences when trigger strategies are used because a single defection signal causes an immediate collapse of cooperation forever. In this section, we present the folk theorem with imperfect public signals (Fudenberg et al.,
Cooperative Homo economicus (1994), which proposes more flexible strategies that are capable of recovering cooperation after a player unintentionally failed to produce the benefit $b$ or he did produce the benefit but was incorrectly perceived to have defected. For instance, a hunter may return with no game either because he had bad luck or because he spent the day sleeping under a tree. The signal is imperfect in the sense that it cannot distinguish between the cases of intentional and accidental defection (execution error) or incorrectly reports the individual’s behavior (perception error).

We will assume that the total expected benefit produced if all $n$ group members cooperate is $nb$, but the actual realized benefit is subject to random variations, so group members cannot infer how many members cooperated by observing the actual value of benefit produced in the period. However, each member receives an imperfect public signal with perceptual error that indicates a defect action accurately, but when the member cooperated, indicates a defection with probability $\epsilon > 0$. Full cooperation entails an expected payoff of $b - c$ for each player. The folk theorem in this case asserts that, no matter how large the group and however closely we wish to approximate the full cooperation payoff, there is a set of incentives that achieves this degree of cooperation, provided $\delta$ is sufficiently close to unity. Perhaps the most striking aspect of the folk theorem with imperfect public signals is that it is true independent of group size and no matter how imperfect the signal, as long as it has a positive correlation with behavior.

In the previous chapter we found that the $n$-player reciprocal altruism model can sustain a high level of cooperation only in very small groups. The folk theorem model thus achieves a degree of efficiency that far outstrips the reciprocal altruism model. This superior efficiency, as we shall see, derives from the assumption that punishment can be directed towards the miscreant rather than all players suffering equally through a breakdown in cooperation.

Suppose, then, that a player who is observed to have defected is punished an amount equivalent to losing $p$ in the current period, the punishment being administered by other members of the group, and the costs of punishment being shared equally among the punishers. We assume that the signal indicating that a player carried out his share of the punishment is public and imperfect, indicating with probability $\epsilon$ that he failed to punish when in fact he carried out the punishment. If a player is signaled as having failed to punish, he is punished by the others in the same manner and degree as a defector is punished. This model has the property that if $A$ is signaled as
having failed to punish B, then B is required to participate in the punish-
ment of A. Moreover, if A and B are both miscreants, each must participate
in the punishment of the other. This arrangement may appear bizarre, but
there are obvious real-world examples, such as a convicted felon paying in-
come taxes that support the costs of his incarceration. Moreover, we could
change this assumption without affecting the conclusion, at the expense of
a more complicated argument.

Suppose the cost of punishing at level \( p \) is \( \alpha p \), where \( \alpha > 0 \). Because
perception errors occur with probability \( \epsilon \) per individual per period, in a
fully cooperative equilibrium, each player will observe an expected number
\( (n - 1)\epsilon \) of defection signals from the other players, which will induce a
total punishment of \( (n - 1)\epsilon p \) of the defectors. Because each punishment is
shared among the other \( n - 1 \) players, each member each period will punish
others an expected amount \( \epsilon p(n - 1)/(n - 1) = \epsilon p \), at cost \( \alpha \epsilon p \).

For cooperation to be a best response, it must be the case that a player
cannot gain by intentionally defecting for one period and then returning
to cooperation in all future periods. If \( \pi_c \) is the single period payoff to
of cooperating assuming all other players cooperate, and \( \pi_d \) is the single-
period payoff to defecting when all others cooperate, followed by returning
to cooperation in the succeeding period, then to maintain cooperation, we
must have \( \pi_c \geq \pi_d \) assuming, as we do, that if it payoffs are identical, the
individual cooperates. We have

\[
\pi_c = b - c - \alpha \epsilon p - \epsilon p
\]

\[
\pi_d = b - (\alpha \epsilon + 1) p
\]

where \( \alpha p \) is the individual’s expected share of the total cost of punishing in
a single period. We can economize on punishment costs by choosing \( p \) to
satisfy \( \pi_c = \pi_d \), which then gives

\[
p = \frac{c}{1 - \epsilon}.
\]

The payoff to each player will then be

\[
\pi_c = b - c \left( 1 + \epsilon \frac{1 + \alpha}{1 - \epsilon} \right),
\]

where the second term in the brackets on the right-hand side is the extra
cost of punishing due to errors. Note the error-induced inefficiency, \( \epsilon c (1 +

PUPSR © Copyright 2011 Princeton University Press.
No part of this book may be posted, distributed, or reproduced in any form or by
any digital or mechanical means without prior written permission of the publisher.
CATALOG
α)/(1 − ε), goes to zero for infrequent errors, independent of group size. Thus, directed punishment appears to solve the problem of cooperation in large groups.

There is a catch, however. Because punishing costs the punisher, self-regarding players have no incentive to carry out the punishment. Thus, this model does not work assuming self-regarding players. We will see in Chapter 5 that people avidly punish defectors even at a cost to themselves, providing a solution to this problem, but one that is based on social preferences rather than self-interest. Equation 4.4 shows that were social preferences admitted in this model the cost to such an altruistic punisher, aec/(1 − e), would be quite small for plausible parameter values. To save the folk theorem, we must thus implement some mechanism to ensure that self-regarding players have an incentive to actually carry out the punishment dictated by the equilibrium. There are potentially many ways to accomplish this, but unfortunately, the proof of the folk theorem does not specify any particular incentive mechanism. This is because the proof is not constructive (Fudenberg et al. 1994). That is, the authors show that an equilibrium with the desired properties exists, but they do not show how to construct one. As a result, in order to explore whether a plausible system of punishment that is incentive compatible for self-regarding agents can be devised, we must construct our own model. This will not be hard to do, and we shall develop the most reasonable model we can devise, but in assessing its plausibility, we must keep in mind there could be some incentive structure of which we are unaware that does not share the limitations of our construction. Here is the best we could devise.

If all individuals cooperate and punish, the expected number of signaled defections will be en. Because n − 1 players punish each defection, there will be en(n − 1) punishments of defectors, which will then generate another e2n(n − 1) defection signals (signals of failure to punish). These new signals of failure to punish engender e2n2(n − 1) new punishment events. The total number of punishments engendered by one period’s failure to cooperate signals is then

\[ \epsilon n(n - 1) + \epsilon^2 n^2(n - 1) + \epsilon^3 n^3(n - 1) + \ldots = (n - 1) \frac{\epsilon n}{1 - \epsilon} \]

Because each individual punishment is p/(n − 1), the total amount of punishment per individual per period will then be

\[ \frac{p\epsilon n}{1 - \epsilon n} \]

(4.5)
provided $\epsilon < 1/n$. We then can write, given (4.3),

$$\pi_c = b - c \left(1 + \epsilon n \frac{1 + \alpha}{(1 - \epsilon n)(1 - \epsilon)}\right).$$  

(4.6)

Assuming $\epsilon$ is considerably smaller than $1/n$, the payoff in this equilibrium is reasonably close to the highest possible equal payoff $b - c$. This is the major message of the folk theorem, although the theorem asserts that there is an incentive structure, doubtless more complicated than ours, that makes the fraction in (4.6) as small as desired for $\delta$ near unity. Note that $\delta$ does not appear in our formulation. This is because we assume that the punishment of an infraction is confined to a single period, which is not the general case. In particular, for $\epsilon n > 1$, a more complex argument must be devised. If we confine our attention to groups no larger than about fifteen and error rates on the order of about 5%, the condition $\epsilon n < 1$ will indeed hold.

It is easy to check that for plausible parameter values (we use $\epsilon = 5\%$, $c = 1$, and $\alpha = 1$), the maximum groups size with positive payoff is $n = 11$ for $b = 4$, $n = 9$ for $b = 3$, and $n = 6$ for $b = 2$. Moreover, the efficiency of the equilibrium, measured by $\pi_c/(b - c)$ fall below 50% for $n = 8$ when $b = 4$, for $n = 6$ when $b = 3$, and $n = 4$ when $b = 2$. Of course if it is less costly to punish than to be punished ($\alpha < 1$), larger group sizes can be supported while maintaining a significant level of efficiency. For instance, for $\alpha = 1/3$ (that is, it costs the punisher one-third to administer a unit of punishment), when $b = 4$, $b = 3$, and $b = 2$, respectively, we can maintain 50% efficiency with groups of size ten, eight, and five, respectively. But even this modest piece of good news evaporates if, as is generally the case, information is private.

### 4.3 The Folk Theorem with Private Information

The case of private signals is more complicated because players cannot co-ordinate their behavior when a defection is detected. With private signals, some players perceive a defection while others perceive cooperation on the part of the same player, and players do not know what perceptions other players may have formed. If defectors are punished by the withdrawal of cooperation in retaliation, Alice cannot tell if Bob defected to punish an observed defection of player Carole, or because Bob defected intentionally, or even that Bob cooperated but Alice received a defect signal from Bob.
by mistake. For example, suppose Bob and Alice repeatedly play the pris-
one’s dilemma depicted in Figure 4.2 with continuation probability $\delta$. We
assume the payoffs to Alice and Bob are distributed only at the termination
of the game, so Alice cannot use current payoffs to determine whether Bob
defected, and vice-versa. However, suppose Alice receives a signal that
informs her correctly when Bob defects, but with probability $\epsilon > 0$ tells
her Bob defected when in fact he cooperated. Suppose Bob has a similar
signal informing him of Alice’s behavior. To see why private signals cause
such a problem, we’ll first see how easy it will be if the two signals were
public, so Bob always knows when Alice receives a defect signal and Alice
always knows when Bob receives a defect signal. Suppose Alice adopts
the “one bad signal triggers defect” (OBSTD) strategy of cooperating on all
rounds until the game ends or she encounters a defect signal, upon which
she defects unconditionally on all future rounds. If Bob adopts the same
strategy, we indeed have a Nash equilibrium because, when Alice receives
a defect signal, she knows that Bob received the same signal and hence will
defect thereafter, the best response to which for Alice is to defect forever
as well. This is, of course, the same model as in §4.1, except now errors
are possible. We show in Appendix A5 that the resulting Nash equilibrium
is fairly efficiency when the error rate is small. For instance, with $\delta = 0.9$
and $\epsilon = 0.05$, the expected payoff if both players could precommit to co-
operating unconditionally is about 45, while the actual payoff to the Nash
equilibrium is about 25, so the efficiency of the Nash equilibrium is about
56%.

Suppose, however, the error signal is private. That is, if Bob cooperates
but Alice receives defect signal, Bob does not know that this is the case.
One might think that this would make little difference because the OBSTD
strategy gives rise to almost the same payoffs. This is indeed the case, as we
show in Appendix A5. The problem is that the pair of OBSTD strategies
is not a Nash equilibrium of the repeated game! In fact, if Bob plays the
OBSTD strategy, Alice’s does better to play the “two bad signals trigger
defect” (TBSTD) strategy. We show in Appendix A5 that for low error
rates, TBSTD has a higher payoff for Alice (and for Bob as well) when
played against OBSTD. The original pair of OBSTD strategies thus did not
form a Nash equilibrium.

While Alice’s TBSTD strategy is better for both Bob and Alice than her
OBSTD strategy when played against Bob’s OBSTD strategy, TBSTD is
not part of a Nash equilibrium. Indeed, if Alice plays TBSTD, then Bob
does better by defecting on purpose every other round, because this never induces Alice to defect until she receives a defect signal when Bob actually cooperated. What, then, is a Nash equilibrium with private signals?

Two different sorts of models for dealing with the private signaling case have been offered. The first, developed in Sekiguchi (1997) and Bhaskar and Obara (2002), players cooperate with probability less than one on the first round and then in following rounds, they use a trigger strategy in which the first time a player receives a defect signal, he defects forever. To explain the first period, and assess the efficiency properties of the equilibrium, assume that from the second period on, Bob and Alice play the prisoner’s dilemma depicted in §4.1, which they play repeatedly with continuation probability \( \delta \). We show in Appendix A5 that the efficiency of the resulting equilibrium is only about 3%.

There is another approach to constructing Nash equilibria for repeated games with private signals (Piccione 2002, Ely and Välimäki 2002, Mailath and Morris 2006). We will not describe this approach because the resulting Nash equilibria require that players use strictly mixed strategies (that is, they randomize over various actions rather than relying on a single action) in each period, and there is no reason for players to actually use such mixed strategies, or to conjecture that the other players will use such mixed strategies. We explain why this is so in the next section.

### 4.4 Dynamically Irrelevant Equilibria

The folk theorem provides no reason to believe that players would ever coordinate on one of the equilibria whose existence the theorem demonstrates, or that should this occur by chance the equilibrium would long survive. This is why we say that the resulting equilibria are dynamically irrelevant. Nor has there been progress towards supplying a realistic dynamical adjustment process that would render such an equilibrium dynamically relevant. It is plausible that individuals are predisposed and able to best respond given full knowledge of the game and the choices of the other players, but when knowledge is imperfect or private and the choices of the other players are not known, surely the most plausible of assumptions, players will not implement the Nash equilibria whose existence is guaranteed by the folk theorems.

Recent research suggests that the conditions for achieving Nash equilibrium are quite stringent and rarely satisfied, except in the simplest of cases
Cooperative Homo economicus (Aumann and Brandenburger 1995). The problem with achieving a Nash equilibrium is that individuals may have heterogeneous and incompatible beliefs concerning how other players will behave, and indeed what other players believe concerning one’s own behavior. Therefore, individuals may choose best responses to strategies that the other players in fact are not playing, resulting in game play that is far from any Nash equilibrium. It is clear from this research that the conditions required for players to implement a Nash equilibrium in all but the simplest games cannot be deduced from the assumption that the players are rational; i.e., that they choose best responses.

In addition to this general point, if cooperation involves the use of mixed strategies, which is the usual case in repeated game equilibria, then players have no incentive to play their equilibrium strategies at all. A for a player is a probabilistic combination of pure strategies, such as playing head or tails each with 50% probability in a pennies-matching game. In a Nash equilibrium in mixed strategies, all the pure strategies a player uses with positive probability must have the same payoff against the mixed strategy choices of the other players, or else the player could increase his payoff by dropping a relatively low-payoff strategy. But then it follows that the player has no incentive to play the Nash best response as opposed to any other combination of the strategies used in the Nash best response. Moreover, the other players know this, and thus have no reason to expect a player to choose the Nash best response, and hence have no reason themselves even to restrict the pure strategies they use to those appearing in the Nash equilibrium.

The founders of game theory were poignantly aware of this problem. John Harsanyi, who was awarded the Nobel prize in economics at the same time as John Nash, offered a very creative resolution in his so-called Purification Theorem (Harsanyi 1973). John Harsanyi showed that in a game with a mixed strategy equilibrium, if there are small random errors in the payoffs, and if a certain technical condition holds, the players will play pure strategies with probabilities approximating the equilibrium mixed strategies. This theorem applies to a number of simple games, but they do not include the repeated games of the form used in the folk theorems, because the required technical condition does not hold in such games (Bhaskar et al. 2004). The upshot of this argument is that the Nash equilibria ensuring efficient cooperation in the public goods game and other social dilemmas are dynamically irrelevant. While there are several models in which individuals“learn”
over time to play a Nash equilibrium (Fudenberg and Levine 1997, Young 2006), the conditions under which these models apply are limited to simple one-shot games, and hence do not apply to repeated games.

This is not to say that the Nash equilibrium concept is irrelevant in practice. In evolutionary game theory we find that every stable equilibrium of a dynamical system governed by a monotone dynamic (one in which higher payoff strategies have higher growth rates than lower payoff strategies), such as the replicator dynamic (Taylor and Jonker 1978), is a Nash equilibrium of the underlying game (Nachbar 1990, Samuelson and Zhang 1992). Here, the analysis runs from the construction of a dynamical system to the analysis of the stable equilibria of such a system, which then must be a Nash equilibrium of the stage game. There is no valid inference in the reverse direction: a Nash equilibrium need not be stable in any evolutionary dynamical setting.

Moreover if, as is surely the case in all real world settings, some signals are private, then trigger and other strategies that require sensitively coordinated responses to defections no longer work because players no longer agree on what happened in the past, and hence cannot coordinate their behavior based on a common understanding of the history of the game. Even in this case, an equilibrium at or near the Pareto frontier for sufficiently patient players can be shown to exist under some conditions (Sekiguchi 1997, Piccione 2002, Ely and Välimäki 2002, Bhaskar and Obara 2002, Mailath and Morris 2006), but we have seen that these equilibria are not very satisfactory. Indeed, if there is a private signaling equilibrium, individuals have no particular incentive to play the strategies that implement the equilibrium, because many other strategies have the same payoffs as the equilibrium payoffs. Moreover, as we have seen, the equilibrium exists only if private signals are very close to being public, so all individuals receive nearly the same signal concerning the behavior of any given group member. When this is not the case, the equilibrium will not exist. Thus, these models apply only to forms of cooperation where all members observe the actions of (nearly) all others with a high level of accuracy.

4.5 Social Norms and Correlated Equilibria

There is an alternative game-theoretic equilibrium concept that does not share the weaknesses associated with the Nash equilibrium described above: the correlated equilibrium. A correlating device is something that sends out
signals, private or public, to the players of a game, indicating which pure strategy each should play. A correlated equilibrium is a situation in which there is a correlating device such that, if all players follow the advice of the correlating device, no player can do better by switching to an alternative strategy.

General arguments support the notion that correlated equilibrium rather than Nash equilibrium is the appropriate equilibrium concept for game theory (Aumann 1987, Vanderschraaf 2001, Gintis 2009a). Assuming players have common knowledge of the game, its rules, and its payoffs, as well as a common belief concerning the probability of the natural events (the so-called moves by Nature) associated with the game, the strategies chosen by rational individuals can then always be modeled as a correlated equilibrium with an appropriate correlating device.

The notion of a correlating device is quite abstract, but one form of correlating device is well known and performs precisely the social function of signaling actions to individuals that, when followed, may lead to a socially efficient outcome. This device is the social norm which, like the choreographer in a ballet, is instituted to issue precise instructions that, when followed, produce the desired outcome. For instance, the system of traffic lights in a city’s highway system instruct drivers when to stop and when to go, and it is normally in the interest of drivers to obey these signals as long as others do so, to avoid accidents.

For a more pertinent example, suppose Alice and Bob repeatedly play the prisoner’s dilemma in Figure 4.2. As we have seen, for a sufficiently large continuation probability, a pair of average payoffs anywhere in the quadrilateral $OABC$ of Figure 4.3 can be achieved as a Nash equilibrium. A choreographer could implement a point on the boundary $AB$ by the judicious choice of instructions to the players. For instance, to achieve the point on $AB$ 1/3 third the distance to $A$, the choreographer could instruct the players to play $(5,5)$ with probability 5/12 and $(−3,8)$ with probability 7/12. This gives Alice the average payoff $5/3$ and Bob the average payoff $25/4$, which is indeed the desired result. Now of course there is no reason for the players to obey the Choreographer in a one-shot game, but if the game is repeated, using the logic of the Folk Theorem where the Choreographer promises to instruct both players to defect forever should one violate the Choreographer’s orders.

It might be difficult to construct a social norm that implements any point on the boundary $ABC$ in Figure 4.3, but perhaps dividing the calendar into
twelfths and assigning five of the resulting days to (5, 5) and the remainder to (−3, 8) might be plausible.

A cooperative equilibrium supported by social norms is one in which not only is the equilibrium strategy mix evolutionarily stable, but also a set of social norms are themselves an evolutionary adaptation, stable against invasion by competing social norms. Sociologists (Durkheim 1933 [1902], Parsons and Shils 1951) and anthropologists (Benedict 1934, Boyd and Richerson 1985, Brown 1991) have found that virtually every society has such social norms, and that they are key to understanding strategic interaction. Borrowing a page from sociological theory, we posit that groups have social norms specifying how a game ought to be played and that these norms are identified as social norms by group members. Learning a social norm includes learning that the norm is common knowledge among those who know it, learning what behavior is suggested by the norm, and learning that a large fraction of group members know the norm and follow it.

Employing the terminology of interactive epistemology (Aumann and Brandenburger 1995), a social norm leads agents to alter their Bayesian priors, and generates a correlated equilibrium with the potential to coordinate cooperative activity and provide incentives for individuals to play their part in this activity (Gintis 2009a).

Social norms do not ensure equilibrium, because error, mutation, migration, deliberate violation of the norm, and other dynamical forces may lead individuals to reject beliefs or behavior fostered by the norm. This may occur because the beliefs might conflict with an individual’s personal experience, or its suggested behavior may be rejected as not in the individual’s best interest; i.e., the action fostered by a social norm must be a best response to the behaviors of the other group members, given the beliefs engendered by the social norm and the individual’s updating. Moreover, social norms cannot be introduced as a deus ex machina, as if laid down by a centralized authority, without violating the objective to provide a “bottom up” theory of cooperation that does not presuppose preexisting institutional forms of cooperation. Social norm are thus discretionary, because any institution that is posited to enforce behavior should itself be modeled within the dynamical system, unless plausible reasons are given for taking a macro-level institution as unproblematically given. Nor are social norms fixed in stone. A group’s social norms are themselves subject to change, those groups producing better outcomes for their members sometimes but not always displacing groups with less effective social norms, and changing social and
demographic conditions leading to the evolutionary transformation of social norms within groups.

4.6 The Missing Choreographer

The economic theory of cooperation based on repeated games proves the existence of equilibria with socially desirable properties, while leaving the question of how such equilibria are achieved as an afterthought. The theory thus exhibits a lack of attention to out-of-equilibrium behavior. This lack of interest in stability and convergence properties renders these models unpersuasive. The folk theorem shares this defect with the even more celebrated Fundamental (“Invisible Hand”) Theorem mentioned at the beginning of the chapter. It purports to model decentralized market interactions, but on close inspection requires an extraordinary level of coordination that is not explained, but rather is posited as a *deus ex machina* (Kirman 1989, Ingrao and Israel 1990, Gintis 2007a).

We have shown, similarly, for the case of cooperation supported by retaliation as in the Folk Theorem, that highly choreographed coordination on complex strategies capable of deterring defection are supposed to materialize quite without the need for a choreographer. As in the case of the Fundamental Theorem, the dynamics are thus unspecified, and, if we are correct, impossible to provide without a fundamental change in the underlying theory.

The failure of the models underlying both the Folk Theorem and the Fundamental Theorem is hardly surprising, for the task we set for them, that of explaining the stability of cooperation among large numbers of self-regarding strangers without recourse to pre-existing norms and cooperative institutions, is not only formidable, it most likely never occurred in the history or prehistory of our species. Humans are indeed exceptional among living creatures in the degree and range of cooperation among large numbers of substantially unrelated individuals. The global division of labor and exchange, the modern democratic welfare state, and contemporary warfare alike evidence our distinctiveness. These forms of cooperation emerged historically and are today sustained as a result of the interplay of self-regarding and social preferences operating under the influence of group-level institutions of governance and socialization that favor Cooperators, in part by helping to coordinate their actions so as to target transgressions for punishment and thus protecting them from exploitation by defectors.
The norms and institutions that have accomplished this evolved over millennia through trial and error. Consider how real world institutions addressed two of the shoals on which the economic models foundered.

First, the private nature of information, as we have seen, makes it virtually impossible to coordinate the targeted punishment of miscreants. In many hunter-gatherer societies the relevant information that would in other societies be private is rendered public by such cooperative customs as eating in public so that violations of sharing norms can be easily detected. Cooperative Japanese shrimp fishermen who pool income across boats deliberately land their catch at an appointed time of day for the same reason (Platteau and Seki 2001). But in most modern societies, where larger numbers are involved, converting private information about transgressions to public information that can provide the basis of punishment often involves civil or criminal trials, elaborate processes that have evolved over centuries and that rely on commonly agreed upon rules of evidence and ethical norms of appropriate behavior. Even with the benefit of these pre-existing social preferences, these complex institutions frequently fail to transform the private protestations of innocence and guilt into common knowledge.

Second, here and in the previous chapter we have seen that cooperation often unravels when the withdrawal of cooperation by the civic-minded intending to punish a defector is mistaken by others as itself a violation of a cooperative norm, inviting a spiral of further defections. We will see evidence from behavioral experiments of this problem in the next chapter. In virtually all surviving societies with substantial populations, this problem is addressed by the creation of a corps of specialists entrusted with carrying out the more severe of society’s punishments. Their uniforms convey the civic purpose of the punishments they mete out, and their professional norms, it is hoped, ensure that the power to punish is not used for personal gain. Like court proceedings this institution works imperfectly.

Both solutions to the problems of cooperation among self-regarding individuals that we have identified, court proceedings for converting private to public information and specialized professional forces to legitimize the punishments that maintain order, presuppose that jurists, police officers, witnesses and others are committed to upholding other-regarding or ethical standards of behavior. It is difficult to imagine that these institutions could work in the absence of these commitments. The economic models of cooperation that presume solutions to these problems thus do not accomplish
their goal, namely explaining cooperation among self-interested individuals.

Modeling the complex processes that sustain human cooperation is a major challenge of contemporary science. Economic theory, favoring parsimony over realism, has instead sought to explain cooperation without reference to social preferences and with a minimalist or fictive description of social institutions. This research trajectory, as we have seen, has produced significant insights. But it may have run its course.

The biological and economic models based on self-regarding individuals thus fail to explain the distinguishing characteristics of human cooperation, namely that it occurs among large numbers of individuals not related by recent common descent. A more plausible explanation takes account of human predispositions to act in ethical, generous, civic minded and also spiteful ways. While the importance of these moral and other-regarding motives has long been doubted by both biologists and economists, recent experiments indicate quite the opposite.
5

Social Preferences

“Is there one word which may serve as a rule of practice all one’s life?”
The Master said, “Is not reciprocity such a word?”

Confucius The Analects. (1938[4th C.B.P.])

A man ought to be a friend to his friend and repay gift with gift. People
should meet smiles with smiles and lies with treachery. A man ought to
be a friend to his friend and also to his friend’s friend. But no one should
be friendly with a friend of his foe.

The Edda (1923[13th C.]), Verses 42,43

5.1 The Structure of Human Morality

Biological and economic models of cooperation based on repeated interac-
tions and reputation formation set themselves an impossible task, namely
explaining the observed propensity of people, even strangers, to contribute
joint projects for common benefit without invoking cooperative norms or
ethical reasons for this behavior. Cooperation is common in large part be-
cause people are motivated by social preferences: they care about the well-
being of others, and value fairness and other norms of decent behavior. Our
explanation is that these social preferences are the proximate cause of al-
truistic cooperation. In this chapter we establish that these preferences are
indeed common.

Consider the ultimatum game (Güth et al. 1982). This is a one-shot,
amerous game in which one subject, called the “proposer,” is given a
sum of money, say $10, and is instructed to offer any number of dollars,
from $0 to $10, to a second subject, called the “responder.” The responder,
who knows how much the proposer was given, can either accept the offer or
reject it. If the responder accepts the offer, the money is shared according
to the offer. If the responder rejects the offer, both players receive nothing,
and the game ends. Because the game is one-shot and anonymous, a self-regarding responder will accept any positive amount of money. Knowing this, by the self-interest axiom just introduced, a self-regarding proposer will offer $1, and this will be accepted.

However, when actually played, the self-regarding outcome is almost never observed and rarely is even approximated. In many replications of this experiment in over thirty countries, under varying conditions and with varying amounts of money at stake, proposers routinely offer respondents very substantial amounts, 50% of the total generally being the modal offer. Respondents frequently reject offers below 25% (Roth et al. 1991, Camerer and Thaler 1995, Camerer 2003, Oosterbeek et al. 2004).

When asked, in post game de-briefings, responders who have rejected low offers often express anger at the proposer’s greed and a desire to penalize unfair behavior. The fact that positive offers are commonly rejected shows that respondents have other-regarding preferences, and the fact that most proposers offer between 40% and 59% of the pie shows that either proposers have other-regarding preferences, or they at least believe respondents have other-regarding preferences, which would motivate them to reject lower offers. Of special interest are the rejections of positive offers. The explanation most consistent with the data is that they are motivated by a desire to punish the proposer for being unfair, even though it means giving up some money to do so. While initially considered odd, these and other experimental results violating the self-interest axiom are now commonplace. Here we survey ten key findings from recent experiments.

Understanding the nature, extent and consequences of social preferences, such as anger at an unfair in a laboratory experiment offer, requires that we consider what behavioral experiments can and cannot show.

5.2 Preferences in the Laboratory

Laboratory experiments add critical information to our understanding of human behavior because, as we will see presently, controlled environments and the experimenter’s ability to manipulate the relevant incentives allow us to distinguish between subtly different hypotheses about preferences. Camerer (2003) and Falk and Heckman (2009) provide overviews of the field. Figure 5.1 lists the experimental games reviewed here and the pages on which the game is described.
Chapter 5

<table>
<thead>
<tr>
<th>Game</th>
<th>Where Described</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Goods Game</td>
<td>§3.3</td>
<td>Yamagishi (1986), Ostrom et al. (1992)</td>
</tr>
<tr>
<td>Ultimatum Game</td>
<td>§5.1, §5.10</td>
<td>Güth et al. (1982), Henrich (2000)</td>
</tr>
<tr>
<td>Gift Exchange Game</td>
<td>§3.2</td>
<td>Akerlof (1982), Fehr et al. (1993)</td>
</tr>
<tr>
<td>Dictator Game</td>
<td>§5.2</td>
<td>Kahneman et al. (1986), List (2007)</td>
</tr>
<tr>
<td>Trust Game</td>
<td>§5.11</td>
<td>Berg et al. (1995), Burks et al. (2003)</td>
</tr>
<tr>
<td>Public Goods, Punishment</td>
<td>§3.3</td>
<td>Fehr and Gächter (2000, 2002)</td>
</tr>
<tr>
<td>Third Party Punishment</td>
<td>§5.5</td>
<td>Fehr and Fischbacher (2004)</td>
</tr>
</tbody>
</table>

Figure 5.1. **Experimental Games.**

Of course, experimental results in the laboratory would not be very interesting if they did not reflect real-life behavior. There is some evidence that they do. Binswanger (1980) and Binswanger and Sillers (1983) used experimental risk-taking lotteries with real financial rewards to predict quite accurately the investment decisions of farmers. Glaeser et al. (2000) explored whether experimental subjects who trusted others in what is called the trust game (described in §5.11) also behaved in a trusting manner with their own personal belongings. The authors found that experimental behavior was a quite good predictor of behavior outside the lab, while the usual measures of trust, based on survey questions, provided virtually no information.

Similarly, Karlan (2005) used a trust game and a public goods game, which are described in §3.3, to predict the probability that loans by a Peruvian microfinance lender would be repaid (the public goods game is described in §6.4). He found that individuals who were “trustworthy” in the trust game were less likely to default. Also, Ashraf et al. (2006) studied Phillipino women, identifying through a baseline survey those women who exhibited a lower discount rate for future relative to current tradeoffs. These women were indeed significantly more likely to open a savings account, and after twelve months, average savings balances increased by 81 percentage points for those clients assigned to a treatment group based on their laboratory performance, relative to those assigned to the control group. In a
similar vein, Fehr and Goette (2007) found that in a group of bicycle messenger workers in Zürich, those who exhibited loss aversion in a laboratory experiment exploring the subjects’ preferences over lotteries also exhibited loss aversion when faced with real-life wage rate changes, and Meier and Sprenger (2010) found that people who display present bias in a laboratory setting are more likely to assume credit card debts.

Moreover, Jeffrey Carpenter and Erika Seki found that Japanese shrimp fishermen who contributed more in a public goods experiment were more likely to be members of cooperatives that shared costs and catch among many boats than to fish under the usual private boat arrangements (Carpenter and Seki 2010). A similar pattern was found among fishermen in the Brazilian northeast, where some fish offshore in large crews whose success depends on cooperation and coordination, while those exploiting inland waters fish singly. The ocean fishers were significantly more generous in a public goods game, ultimatum game and dictator game than the inland fishers (Leibbrandt et al. 2010).

A better test of the external validity of experiments would include a behavior based measure of how cooperative the individuals were, not simply whether they took part in a cooperation-sensitive production process. The Brazilian fishers provide just such a test. Shrimp are caught in large plastic bucket-like contraptions. Holes are cut in the bottom of the traps to allow the immature shrimp to escape, thereby preserving the stock for future catches. Smaller holes catch more shrimp but compromise the future stock. The fishermen thus face a real world: the present value of expected income of any given shrimper would be greatest if they cut only small holes in their own traps while others cut large holes in theirs. Small trap holes are a form of defection, and just as in the public goods game the small trap hole defection is the dominant strategy for a self-regarding individual: whatever the others do a self-regarding shrimper would cut small holes. But a shrimper might resist the temptation to defect if he were both public spirited towards the other fishers and sufficiently patient to value the future lost opportunities that larger holes would entail. Fehr and Leibbrandt implemented both a public goods game and an experimental measure of impatience with the shrimpers. They found that both patience and cooperativeness in the public goods game predicted significantly larger trap holes (Fehr and Leibbrandt 2010). The effects, controlling for a large number of other possible influences on hole size, are substantial. A shrimper whose experimentally measured patience and cooperativeness is a standard deviation greater than the
mean is predicted to cut holes in his traps that are half a standard deviation larger than the mean.

Additional evidence of external validity of comes from a set of experiments and field studies with 49 groups of herders of the Bale Oromo people in Ethiopia who were engaged in forest commons management. Devesh Rustagi and his co-authors implemented public goods experiments with a total of 679 herders. They also studied the success of the herders’ cooperative forest projects. The most common behavioral type in the experiments, constituting a bit more than a third of the subjects, were “conditional cooperators” who responded positively to higher contributions by others. Controlling for a large number of other influences on the success of the forest projects the authors found that groups with more conditional cooperators were significantly more successful, in terms of number of new trees planted, than groups with fewer reciprocators. This was in part because members of groups with more conditional cooperators spent significantly more time monitoring the use of the forest by others. As in the case of the Brazilian shrimpers, the effects of group composition were large. A ten percent increase in the fraction of experimentally identified conditional cooperators in a group was associated with an increase in trees planted or time spent monitoring of about a 3 percent (Rustagi et al. 2010).

But these examples of correspondence between experimental and real-world behavior are hardly decisive. Consider, for example, the dictator game in which a one subject (the Dictator) is assigned an endowment of money and asked to allocate some portion of it (including none) to a passive recipient. Typically more than 60% of subjects allocate a positive sum to the recipient, which the average is about a fifth of the endowment. We would be sadly mistaken if we inferred from this that 60 percent of individuals would spontaneously transfer funds to anonymous passers-by, or that the same subjects would offer a fifth of the bills in their wallet to a homeless person asking for help. Subjects who reported that they had never given to a charity allocated 60 percent of their endowment to a named charity in a lab experiment (Benz and Meier 2008). Thus there are counter-examples to the above evidence. A more subtle problem with the studies of experimental validity is the low bar that is implicit in most of these tests. Merely a positive, statistically significant, correlation between of experimental play and real behavior is taken as evidence for validity. A more demanding test would ask whether the effect size or covariance is what one would expect from a valid measure.
Most individuals are strongly influenced by the cues of appropriate behavior offered by the situation in which an action is taken (Ross and Nisbett 1991), and there is no reason to think that experiments are immune to this context-dependent aspect of individual behavior (Levitt and List 2007). Validity concerns arise from four aspects of human behavioral experiments that do not arise in most well-designed natural science experiments. First, experimental subjects typically know they are under a researcher’s microscope, possibly inducing different behaviors than would occur under total anonymity or under the scrutiny of neighbors, family or workmates. Second, interactions with other subjects are typically anonymous and without opportunities for ongoing face to face communication, unlike many social interactions. Third, subject pools may be quite different from the real-world populations of interest, in part due to the process of recruitment and self-selection. Finally, many of the experiments that provide evidence for the salience of social preferences are deliberately structured as strategic interactions like the Ultimatum Game that give scope for ethical or other-regarding behavior that may be absent in competitive markets and other important real world settings.

It is not clear whether these four aspects of behavioral experiments induce the greater expression of social preferences in the lab or the opposite. For example, the fact that in most cases subjects are paid a show up fee to participate in an experiment might attract the more materially-oriented while the fact that most subjects are students may bias the results in other ways. Available evidence, however, suggests that the students volunteering for experiments are not more pro-social than those who do not, and that if anything university students are more likely to be self-regarding than older non students (Fehr and List 2004, Carpenter et al. 2005, Cardenas 2005, Sutter and Kocher 2007, Bellemare et al. 2008, Falk et al. 2010, Burks et al. in press). We report on several studies supporting this conclusion in Appendix A6.

While warranting caution in generalizing the details of experimental behavior to the real world none of these validity concerns is sufficient to dismiss the experimental evidence that social preferences are important behavioral motivations. This especially the case when experiments identify motives that allow a consistent explanation of the otherwise anomalous real world examples of self-sacrifice and generosity in wage setting, voting, and tax compliance mentioned in Chapter 2 (Falk and Heckman 2009). Moreover, studies typically find that older subjects exhibit at least as much
prosociality and strong reciprocity as college students (Sutter and Kocher 2007, Bellemare et al. 2008). In the penultimate section of the chapter, we will reconsider to these issues in light of the experimental results to be presented.

5.3 Strong Reciprocity is Common

Recall that strong reciprocators have a predisposition to cooperate in situations where this is beneficial to others, and they respond to others’ cooperative behavior by continuing or enhancing their level of cooperation, while responding to lack of cooperation by others in these situations (and to violations of ethical norms more generally) by punishing the offenders, even at a material cost to themselves, and even when they cannot expect future personal gain from such behavior. Strong reciprocity is common. In dyadic interactions, people tend to cooperate, and to reward the cooperation of their partners, while punishing the free-riding of their partners, even when they cannot expect to gain from such behavior.

We have already described the basic prisoner’s dilemma game, the key feature of which is that mutual cooperation maximizes the sum of the players’ payoffs, but defecting on one’s partner maximizes a player’s payoffs independently of what the partner does. Thus among individuals with self-regarding preferences, knowledge of the partner’s strategic choice should make no difference in the outcome. The result should be mutual defection. The fact that a high level of cooperation results when subjects are assured that their partner has already cooperated or that their own decision to cooperate would be transmitted to their partners prior to the latter choosing what to do is an indication that subjects are motivated by strong reciprocity. Knowing that your partner has cooperated changes the subjectively relevant payoffs in the game. One could still make more money by defecting, but the cooperativeness of the partner motivates reciprocation rather than exploitation.

Another experiment suggesting that strong reciprocity is common is the gift exchange game, modeled as the “experimental labor market” investigated by Fehr et al. (1997). The authors divided a group of 141 subjects into a set of “employers” and a set of “employees.” If an employer hires an employee and pays wage $w$, with $0 \leq w \leq 100$, his profit is $\pi = 100e - w$, where $0.1 \leq e \leq 1$ is the amount of “effort” exerted by the employee. The payoff to the employee is then $u = w - c(e)$, where $c(e)$ is a “cost of effort”
function that is increasing at an increasing rate (i.e., $c_1, c_2 > 0$). All payoffs involve real money that the subjects are paid at the end of the experimental session.

The employer then offers a “contract” specifying a wage $w$ and a desired amount of effort $e^*$. A contract is made with the first employee who agrees to these terms. An employer can make a contract $(w, e^*)$ with at most one employee. The employee who agrees to these terms receives the wage $w$ and supplies an effort level $e$, which need not equal the contracted effort, $e^*$. In effect, there is no penalty if the employee does not keep his promise, so the employee can choose any effort level with impunity. Although subjects may play this game several times, each employer-employee interaction in a given game is a one-shot (non-repeated) event.

If employees are self-regarding, they will choose the zero-cost effort level, $e = 0.1$, no matter what contract they have agreed to. Knowing this, self-regarding employers will never pay more than the minimum necessary to get the employee to accept a contract, which is 1. The self-regarding employee will accept this offer, and will set $e = 0.1$, giving him payoff $u = 1$. The resulting employer’s payoff is $\pi = 0.1 \times 100 - 1 = 9$.

In fact, however, this outcome rarely occurred in this experiment. Indeed, the higher the employer’s choice of demanded effort, the more both employers and employee’s earned. In effect, employers presumed the strong reciprocity predispositions of the employees, making more generous wage offers and receiving higher effort.

The above evidence is compatible with the notion that the employers were purely self-regarding, because their seemingly generous behavior vis-à-vis their employees was effective in increasing employer profits. To see if employers were also strong reciprocators, following this round of experiments, the experimenters extended the game by allowing the employers to respond to the actual effort choices of their workers: at a cost of 1, an employer could increase or decrease his employee’s payoff by 2.5. If employers were entirely self-regarding, they would of course do neither, because they do not interact with the same worker a second time, so a self-regarding employer would consider punishing a shirker or rewarding a hard worker to be just throwing away money. However, 68% of the time, employers punished employees that did not fulfill their contracts, and 70% of the time, employers rewarded employees who overfulfilled their contracts. Indeed, employers rewarded 44% of employees who exactly fulfilled their contracts. Moreover, employees expected this behavior on the part of their employ-
ers, as shown by the fact that their effort levels increased significantly when their bosses gained the power to punish and reward them. Underfulfilled contracts dropped from 86% to 26% of the exchanges, and overfulfilled contracts rose from 3% to 38% of the total. Finally, allowing employers to reward and punish led to a 40% increase in average net payoffs, even when costs associated with employer punishment of employees are taken into account.

We conclude from this study that the subjects who assume the role of “employee” reciprocate seemingly generous offers by employers, even when they are certain there are no material repercussions from behaving in a self-regarding manner. Moreover, subjects who assume the role of “employer” expect this behavior and make higher payoffs when they take this into account. “Employers” reward good and punish bad behavior when they are permitted to punish, even when their payoffs would be maximized by refraining from rewards and punishment. Finally, “employees” expect employer rewards and punishments, and adjust their own effort levels accordingly.

A large number of additional experiments with the game have replicated these results (Gächter et al. 2004, Fehr et al. 2009).

5.4 Free-riders Cause Cooperation to Unravel

In a social dilemma that is repeated for a number of periods, subjects tend to start out with a positive and significant level of cooperation, but unless there are very few free-riders in the group, cooperation subsequently decays to a very low level.

The experimental public goods game is designed to illuminate such problems as the voluntary payment of taxes and contribution to team and community goals (Ledyard 1995). The following is a common variant of the game. Ten subjects are told that $1 will be deposited in each of their “private accounts” as a reward for participating in each of the ten rounds of the experiment. For every $1 that a subject moves from his “private account” to the “public account” on a given round, the experimenter will add one half dollar to the final payoffs to each of the subjects. At the end the ten rounds, the subjects are given the total of their final payoffs, and the experiment is terminated.

The sum of individual payoffs will be maximized if in each round, each puts $1 in the public account, generating a public pool of $10. In this
case, the experimenter adds $5 to the final payoff of each subject. At the end of the game, ten rounds having been played, each subject would be paid $50—in each round $1 contributed minus the $0.50 returned by the experimenter. However, every $1 a player contributes to the public account, while benefitting the nine others by a total of $4.50, costs the contributor $0.50. Therefore the dominant strategy for a self-regarding player is to contribute nothing to the pool, and if all subjects do this, each then earns just $10. The experimental public goods game is thus an \( n \)-person prisoner’s dilemma (§3.3).

In fact, as in prisoner’s dilemma experiments (Fehr and Fischbacher 2002, Fischbacher and Gächter 2010), in public goods experiments, only a fraction of subjects conform to the self-regarding actor model, contributing nothing to the public account. Rather, subjects contribute on average about half of their private account on round one, but in later rounds, contributions decay to a level close to zero.

This decay result is significant for the following reason. A supporter of the self-regarding actor model is inclined to interpret other-regarding behavior in experiments as confusion on the part of the subjects, who are not accustomed to anonymous interactions. Their behavior therefore reflects their beliefs, not their preferences. In everyday life, one’s actions are normally seen by others, so a failure to contribute would entail a loss of reputation, and hence a loss of future profitable exchanges. The anonymity of the laboratory may be sufficiently extraordinary that subjects simply play by these rules of everyday life. Accordingly, the decline in contributions in the public goods game might be seen as a confirmation of this belief-based interpretation: subjects are learning how to maximize their payoffs through game repetition.

However, were this explanation correct, if the same subjects were permitted to play a second multi-round public goods game identical to the first, they should refuse to contribute on the very first round. Andreoni (1988) and Cookson (2000) tested this prediction, and found it to be wrong. When the public goods game is played with several groups and after every series of rounds group membership is reshuffled and the game is restarted, subjects begin each new series by contributing about half, but each time cooperation decays in the later rounds. If one believes that the decay in contributions within a game is due to learning how to maximize payoffs in the context of anonymity, one would also have to believe that subjects unlearn the money-maximizing behavior between series! In fact, the only
A reasonable explanation for the decay of cooperation is that public-spirited contributors want to retaliate against free-riders, and the only way available to them in the game is by not contributing themselves. Subjects often report this reason for the unraveling of cooperation retrospectively.

Another indication that free-riding and the retaliation against free riders is the cause of the unraveling of cooperation comes from an experiment by Page et al. (2005). The experimenters compared four baseline sessions, each of which included 16 subjects in a 20-round public goods game, with four sessions in which, after three 20-round games, subjects were given a list of the average contributions of the other players in all four groups, and were permitted to rank their preference for playing with one or more of these subjects. Subjects who ranked each other highly were assigned to the same group, and subjects who were not ranked highly by others were also assigned to the same group.

In baseline treatments, contributions began at an average of 60% and declined to 9% in the last round, for an average contribution rate of 38% of the endowment over the twenty rounds. Where subjects could choose their partners, cooperation did not decay over time, and the average contribution rate was 70% of the endowment. Note that this high average cooperation rate includes the performance of low contributors, who were obliged to play with one another.

To understand this result, note that when subjects could choose their partners, there was a strong tendency for subjects to play with others who approximately share their level of contribution. This is because the experimenters would always satisfy the request of two players who preferred to be together before the request of a pair only one of whose members preferred to associate with the other. Thus the top of the four elective groups maintained an average contribution rate of over 90% with no tendency to decay, except for an end-game effect in the last three rounds that brought contributions down to about 60%. The second most preferred group maintained an 80% average, with a similar end-round effect, while the third group averaged about 65%, with a relatively weak tendency to decay, from about 75% in the first rounds to 60% in rounds 12 to 16, and then to about 50% in the final three rounds. The lowest group showed the usual decay from 75% contribution in the first three rounds to 10% in the final round, for an average of 45%. These results are consistent with the idea that the decay of cooperation is due to relatively high contributors reacting to low contributors by lowering their own contribution. When subjects in the same group
are relatively uniform in their contributing behavior, this decay mechanism is attenuated.

These experiment show that when those predisposed to cooperate can associate preferentially with like-minded people, cooperation is not difficult to sustain. We return to this basic rule in the next and subsequent chapters. These and a host of related experiments provide strong support for the notion that cooperation is due to reciprocal preferences, not the refusal of subjects to believe that the game is really anonymous, and hence their behavior cannot have long-term reputational effects.

### 5.5 Altruistic Punishment Sustains Cooperation

In social dilemmas, strong reciprocators, by punishing free-riders, induce their cooperation in subsequent play, thereby allowing cooperation to be sustained over time. Experiments by Orbell, Dawes, and Van de Kragt (1986), Sato (1987), Yamagishi (1988a), (1988b), (1992), and Ostrom, Walker and Gardner (1992) show that when subjects are given a direct way of retaliating against free-riders rather than simply withholding their own cooperation, they use it in a way that helps sustain cooperation. A particularly clear example of this was given by Fehr and Gächter (2000, 2002), who designed a repeated public goods game with an option of costly retaliation against low contributors in some treatments.

Fehr and Gächter (2000a) used a six- and a ten-round public goods game with four-person groups, employing three different methods of assigning members to groups. Under the partner treatment, the four subjects remained in the same group for all ten periods. Under the stranger treatment, the subjects were randomly reassigned after each round. Finally, under the perfect stranger treatment the subjects were randomly reassigned in such a way that they would never meet the same subject more than once, so subjects knew that costly retaliation against low contributors could not possibly confer any pecuniary benefit to those who punish. Subjects were informed which treatment would obtain for their experiment.

Fehr and Gächter ran the experiment for ten rounds with punishment and ten rounds without. Their results are illustrated in Figure 5.2. The experimenters found that subjects were more heavily punished, the more their contributions fell below the average for the group. As a result, when costly punishment was permitted, cooperation did not deteriorate, and in the Partner treatment, despite strict anonymity, cooperation increased to almost full
cooperation, even on the final round. When punishment was not permitted, however, the same subjects experienced the deterioration of cooperation found in previous public goods games.

This result is telling because in the stranger and perfect stranger treatment, punishing is itself a public good, so punishing low contributors is no different from contributing to the public good; both confer benefits on others at a cost to oneself. In both treatments, not contributing and not punishing are dominant strategies (they maximize payoffs irrespective of the actions of the others.) We term punishment in this setting altruistic for this reason. Yet as we saw subjects treat contribution and punishment differently. After the initial rounds in the standard public goods without punishment game, experimental subjects decline to contribute altruistically but once punishment is permitted they avidly engage in the altruistic activity of punishing low contributors.

---

82 Chapter 5

---

Figure 5.2. **Average Contributions over Time.** Partner, Stranger, and Perfect Stranger Treatments are shown when the Punishment Condition is Played First (Fehr and Gächter, 2000). Results are similar when the punishment condition is played second.
Social Preferences

Part of the reason for the difference is that people have an intrinsic motivation to punish shirkers, not simply an instrumental desire to alter their behavior or to affect the distribution of payoffs to either reduce unfairness or to enhance one’s own relative payoffs. This is similar to what Boyd and Richerson (1992) call retribution punishment and the negative analogue of Andreoni’s (1990) warm glow altruism. That subjects view punishment of shirkers also as retribution rather than simply as instrumental towards affecting behavior is consistent with the recent public goods with punishment experiment of Falk et al. (2005). The game was one shot, ruling out behavior modification as a motive for punishing low contributors, and the punishment technology was such that punishment could not alter the difference in payoff between the punisher and the target (the cost to the punisher was the same as that inflicted on the target). Nonetheless, sixty per cent of cooperators punished defectors.

Further evidence for our assumption that punishment is non-strategic comes from the public goods experiment of Fudenberg and Pathak (2009). As in the standard game, following each round of contributions subjects were given information on the contributions of fellow group members and had the opportunity to deduct some of their own payoffs in order to lower the payoffs of another in the group. But unlike the usual treatment, in which the targets of punishment were informed of the level of punishment received after each round, in the Fudenberg and Pathak experiment the levels of punishment were not revealed until the experiment was over, and those who punished others knew this. Thus the experimental design ruled out modifying the behavior of shirkers as a motive for punishment. Consistent with what the authors term a “pure preference” motivation for punishment, subjects nonetheless punished shirkers, leading the authors to conclude that “agents enjoy punishment, where ‘enjoyment’ includes anger and a desire for retribution.” There is considerable further evidence for our non-strategic modeling of punishment (deQuervain et al. 2004).

5.6 Effective Punishment Depends on Legitimacy

Another public goods with punishment experiment confirms that altruistic punishment enhances group-level cooperation. But it raises a new question: do groups that punish free-riders actually benefit or do the costs of punishing outweigh the benefits to cooperation that result? Benedikt Herrmann, Christian Thöni and Simon Gächter (2008) chose an unusually diverse set
of subject pools from fifteen populations ranging from familiar experimental sites such as Boston and Zurich to the less frequently studied Riyadh, Muscat, and Chengdu (China) to implement a ten-period similar to that in the Fehr and Gächter experiment. As expected, cultural differences among the subject pools were significant, but in all of them subjects contributed substantial amounts in the first period, and in the absence of the punishment option, in subsequent periods cooperation unraveled. As in earlier experiments, when the punishment option was available it was widely used, especially in the early periods, and as a result the unraveling of contributions did not occur.

Figure 5.3. Free-rider and Anti-social Punishment in Sixteen Cities. The grey bars refer to punishment of those contributing the same or more than the punisher. The black bars indicate punishment of those contributing less than the punisher.

What surprised the experimenters was the fact that, averaging over the ten periods, most of the subject pools had higher average payoffs when the punishment option was precluded. They readily identified the problem:
in many societies a significant amount of punishment was directed at high contributors, possibly as a retaliation against punishment received in earlier rounds by subjects who believed that it was the high contributor who were doing most of the punishment (Figure 5.3). The result was vendetta-like retaliation against punishment leading to costly arms-race dynamics of wasteful punishment expenditures. The authors termed this “anti-social punishment.” Other experiments have found the same patterns (Cinyabuguma et al. 2006, Ertan et al. 2005, Ertan et al. 2009).

A closer look at the data reveals that all but two of the subject pools improved their net payoffs over time, with about half of them gaining net benefits by the penultimate period. In most subject pools subjects took advantage of the last round to free ride on the contribution stage, expecting, in most cases falsely, it turns out, that their fellow group members would also free ride and not punish in the last period. In most subject pools, it appears, there was an initial learning period in which those predisposed to free ride by not contributing were heavily punished. The result was both low net benefits in early rounds and more cooperative behavior in contributing in subsequent rounds which obviated the occasion for such frequent punishment. However, as is clear in Figure 5.4, the impact of anti-social punishment on average payoffs was very strong.

To test the possibility that the net returns to having a punishment option are high when the game is repeated a sufficient number of periods, Gächter, Elke Renner and Martin Sefton (2008) implemented an identical game with University of Nottingham students, but allowed the subjects to interact over fifty periods. They found, as the learning interpretation suggested above had led them to expect, that after the initial rounds, the net benefits to the group with the punishment option significantly exceeded those of the no-punishment group with the difference in net payoffs growing monotonically over time, except for the final round in which the hapless end-game free riders were heavily punished (Figure 5.5).

Given that most social dilemma interactions, in neighborhoods, work teams, and the like, extend over far more than ten periods we find the concern that altruistic punishment lowers group benefits to be misplaced. The experiment of Dreber et al. (2008) does not constitute evidence for the counter-productive punishment hypothesis for the additional reason that their two-person game made punishment irrelevant, for one could always retaliate on a defector simply by withdrawing cooperation, thus obviating the need for any special kind of punishment. But while the fifty-period de-
Figure 5.4. **Anti-social Punishment Lowers Average Payoffs.** Shown is the group average earnings in the treatment in which punishment was allowed divided by the average earnings in the no-punishment treatment. Excepting the end-game effects (lower contributions, greater punishment) subjects in Chengdu, Boston and Melbourne on average benefitted from the opportunity to punish once they achieved a high level of contribution which then induced little punishment. This did not occur in Athens and Muscat.

The design of the Gächter et al experiment corrects one of the design biases that suggested counter productive punishment in the earlier experiment, their design still misses something essential to altruistic punishment in the real world: it is effective only if it is regarded as legitimate according to widely-held social norms. We model this legitimacy aspect of altruistic punishment in Chapter 10.

Ertan et al. (2009) designed an ingenious experiment to explore this possibility. They allowed experimental subjects prior to playing the public goods game to vote on whether punishment should be allowed and if it should be restricted in any manner. Here is what they found. From their first opportunity to vote, no group ever allowed punishment of high contributors, most groups eventually voted to allow punishment of low contributors, and the result was both high contributions and high efficiency levels. Thus, we
found evolution in the lab of societies that solve their free rider problems by allowing low contributors alone to be punished. Apparently the determination of the punishment system by majority rule made the punishment not only an incentive but also a signal of group norms.

This experiment suggests a possible explanation of the Herrmann et al. (2008) study of fifteen different cultures. Punishment of free riders, even if they were strangers, was legitimate in Boston, Copenhagen, and Chengdu but it was not in Mucat and Athens. We will see in Chapter 7 that ethnographic studies of the punishment of group members is sometimes finely tuned to achieve legitimacy: it is coordinated by gossip and rarely carried out by a single individual. We will also see that in small-scale societies punishment can be highly effective even when it takes the form of ridicule or gossip and it inflicts no material costs on its targets. The importance of the moral signal conveyed by punishment rather than simply the material incentive that it provides is also suggested by experiments.
5.7 Purely Symbolic Punishment is Effective

People are sensitive to others’ evaluation of their moral worth or intentions, and will cooperate in social dilemmas when the punishment for free-riding takes the form of criticism by peers rather than a reduction in material payoffs.

To test this idea, Masclet et al. (2003) allowed the subjects in a public goods game to assign “disapproval points” to the other group members after the subjects have been informed about each others’ contributions. These disapproval points have no material consequences. They merely indicate the members’ evaluation of one another. Disapproval alone raised the contributions to the public good relative to the baseline with no punishment opportunities.

In another experiment, Bochet et al. (2006) compared the usual baseline public goods game with a “chat-room” situation in which four group members communicated with one another through their computer terminals for several minutes before each round, and a “face-to-face” situation in which group members engaged in face-to-face communication. These treatments are often called “cheap talk” by game theorists, because promises made cannot in any way be enforced. Nevertheless, the experimenters found that both forms of communication increased contributions considerably above the baseline level. Surprisingly, chat-room communication was almost as effective in increasing contributions as face-to-face communication, and adding the option of punishing increased contributions little more. Specifically, (a) face-to-face, face-to-face with punishment, chat room, and chat room with punishment all induced average contribution rates above 95%, and about 85% in the last of the ten rounds; (b) punishment alone performed considerably less well, averaging about 70%, and 60% in the last period; (c) the baseline (no communication, no punishment) treatment performed worst of all, starting at 60% cooperation and declining to 20% in the final period, for an average contribution of about 48%.

This is consistent with the results of a public goods with punishment experiment implemented in 18 rural communities in Zimbabwe by Abigail Barr (2001). The game was structured along the above lines, except for the punishment stage, in which there was no option to reduce the payoffs to others. Rather, following the contribution stage, Barr’s assistant would stand beside each player in turn and say to the group as a whole, “Player number __, Mr/Mrs __ contributed ___. Does anyone have anything to say about
that?” A quarter of the participants were criticized for contributing too little (“stingy,” “mean,” “Now I know why I never get offered food when I drop by your house!”). Five percent were criticized for giving too much (“stupid,” “careless with money”). Those who made low contribution and were criticized made larger contributions in subsequent rounds. Moreover, those who contributed a low amount and escaped criticism, but had witnessed the criticism of others who had contributed a similar amount, increased their contributions by even more than those directly criticized. Also, those who had contributed a large amount and were criticized reduced their contribution in subsequent rounds. Where low contributions escaped criticism entirely, contributions fell in subsequent rounds.

Gächter and Fehr (1999) also found that, given some minimal social contact among strangers, making individual contributions publicly observable raises contributions to the public good substantially. Beyond this, Gächter and Fehr asked subjects to fill out questionnaires to measure the strength of their emotional responses towards cooperation and free-riding on the part of others. They show that free-riding elicits extremely strong negative emotions among the other group members. Moreover, in the post-experimental group discussions the other group members verbally insulted the free-riders.

5.8 People Punish Those Who Hurt Others

People punish not only those who have hurt them personally, but also those who violate social norms and hurt others who are situated similarly to themselves.

Fehr and Fischbacher (2004) studied a “third party punishment” game with three players. The game between player A and player B is a “dictator game”, in which Player A gives a certain amount of money to player B, who has no say in the matter. In this experiment, Player A was given an endowment of 100 tokens of which he could transfer any amount to player B (at the end of the game, the tokens are converted into real money). Player C has an endowment of 50 tokens and observes the transfer of player A. After this player C can assign punishment points to player A. For each punishment point assigned to player A, player C has costs of 1 token and player A incurs a penalty of 3 tokens. Because punishment is costly, a self-regarding player C will never punish. However, if there is a sharing norm player C may well punish player A if A gives too little.
In the above experiments player A’s were never punished if they transferred 50 or more tokens to player B. If they transferred less than 50 tokens the punishment was the stronger the less player A transferred. A player A who transferred nothing received on average 9 punishment points from player C, so player A’s payoff was reduced by three times this, or 27 tokens. A selfish player A in this game might still prefer not to give, but with several player C’s, his cumulative punishment might be sufficient to induce even a selfish player A to make an equitable gift to player B.

Engelmann and Fischbacher (2009) studied “indirect reciprocity” and “strategic reputation-building” in an experimental helping game (§3.4). Indirect reciprocity occurs when player C is likely to punish player B when player B has been unfair to player A, and is likely to reward player B when player B has been nice to player A. Strategic reputation-building occurs when player C behaves in the above manner only when his actions are seen by others, and hence can help build a reputation for social behavior. Of course, unless there are indirect reciprocators, strategic reputation-building can have no effect. Nevertheless, it is interesting to see that even self-regarding individuals may engage in third-party punishment if they believe that this will induce other-regarding individuals to behave favorably towards them. In their experiment, at any time only half of the subjects were capable of building a reputation by having their behavior observed by the group. Engelmann and Fischbacher found that, while non-strategic indirect reciprocity appears to be important, helping behavior was influenced at least as much by strategic considerations. Strategic reciprocators did better than pure indirect reciprocators and, of course, selfish types had the highest payoffs of all. This experiments shows that strong reciprocators will punish violators of social norms even when they themselves are not directly hurt by the violator. Punishment is thus not simply retaliation in response to personal damages but appears to reflect more general ethical norms (Ule et al. 2009). Self-regarding types will mimic the behavior of strong reciprocators to receive favorable treatment from them.

5.9 Social Preferences are not Irrational

The desire to contribute, to punish, and otherwise to satisfy social preferences, like the desire for conventional goods and services, can be represented by preferences that conform to standard definitions of rationality (Savage 1954, Hechter and Kanazawa 1997, Gintis 2009a). These prefer-
ence imply observable trade-offs, depending on the costs, and experiments confirm that the higher the cost of moral behavior, the less its frequency.

Many observers of experimental games have interpreted the fact that people sometimes sacrifice material gain in favor of moral sentiment as an indication of irrationality, the term “rationality” being used as a synonym for “consistent pursuit of self-interest.” But, subjects appear to be engaged in the same sort of optimizing when deciding to cooperate and punish as when they compare prices to decide what to cook for dinner. This suggests that the preferences that lie behind their social behavior are consistent with the basic axioms of rationality, namely on transitivity and completeness.

Andreoni and Miller (2002) tested the “rationality” of moral choices by asking 176 subjects to play a version of the dictator game. Recall that in the dictator game, subject A is given a sum of money by the experimenter, and asked to transfer whatever proportion of the money that he wishes to another (anonymous) subject B. After A makes his decision, the money is transferred, and the game is over. In the Andreoni-Miller version of the game, the cost of giving was varied by the experimenter. A is given a sum $m$, a price $p$, and is asked to keep an amount $\pi_s$, while transferring an amount $\pi_o$ to B, such that the budget equation is $\pi_s + p\pi_o = m$. Thus for instance, if $m = 40$ and $p = 3$, A could keep all 40 for himself, or could keep ten and transfer ten to B, thus satisfying the equation $10 + 3 \times 10 = 40$.

This $p$ is the price of generosity. By varying $m$ and $p$, the experimenters could see if the subjects responded to changes in the price of generosity in the expected way, and thus had “rational preferences.”

In this experiment, 75% of the “dictators” gave away some money, showing other-regarding behavior, and the average amount given away was 25.5% when the price $p = 1$ (a dollar-for-dollar transfer), which is about the same as in other dictator games (Forsythe et al. 1994). Moreover, the higher the price of generosity, the less money was given. For instance, when it cost two dollars for each dollar passed to the other person ($p = 2$), only 14.1% was given away, and when it cost four dollars for each dollar passed, only 3.4% of the dictator’s endowment was passed. Finally, only 18 of the 176 subjects violated the principle of transitive preferences that requires that if an individual prefers $A$ over $B$ and $B$ over $C$, he then prefers $A$ over $C$. Moreover, these violations were almost all very minor. Indeed, 98% of the individual choices were consistent with transitive preferences.

Similarly, in a public goods with punishment experiment in which punishment cannot be motivated by self-regarding preferences similar to that
of Fehr and Gächter (2000a), Anderson and Putterman (2006) found that the level of altruistic punishment that subjects inflicted on others varied inversely with the cost of punishing.

The fact that other-regarding preferences support price-responsive behaviors conforms to our representation of social preferences as distinct motivations within the framework of transitive preference rather than some *sui generis* irrational or non-rational mode of behavior (Fisman et al. 2007). The fact that for many experimental subjects virtue is its own reward is perfectly consistent with the fact that, as in the case with those with self-regarding preferences, they would consider the price.

## 5.10 Cultures and Institutions Matter

As experimental evidence has accumulated showing that many, perhaps most, individuals are not entirely self-regarding, it appeared natural to wonder if an alternative equally simple representation of human behavior, but one stressing ethical and other-regarding motives, might be universally valid. As the previous century drew to a close we were impressed by how similar were the experimental results coming from experimental laboratories around the world, from Beijing, Tel Aviv, Pittsburgh, Zurich and Ljubljana. Could *Homo economicus* simply be replaced by *Homo sociologicus*, *Homo altruisticus*, or as we once suggested, *Homo reciprocans*?

It is too early to say, but we doubt that such a universal model will prove viable. The social preferences that become salient in a population depend critically on the manner in which a people’s institutions and livelihood frame social interactions and shape the process of social learning. A result is substantial cross-cultural differences in the nature and extent of social preferences.

We have already discussed the remarkable differences across subject pools from around the world in the public goods with punishment game. Here we report ultimatum game experiments in which the subject pool is not, as is usually the case, university students, but instead were members of fifteen small scale societies with little contact with markets, governments or modern institutions. With our colleagues, a team of seventeen anthropologists and economists, we designed these experiments to explore whether the results reported above are common in societies with quite different cultures and social institutions (Henrich et al. 2001, Henrich et al. 2004). The fifteen societies included hunter-gathers, herders, and farmers.
Our results strongly affirmed cultural difference in experimental play. Among the Au and Gnau people in Papua New Guinea, offers of more than half of the amount provisionally allocated to the proposer were common, and even splits were commonly accepted. Moreover, both high and low offers were rejected with equal frequency. This seemingly odd result is not surprising in light of the practice of competitive gift giving as a means of establishing status and subordinacy in these and many other New Guinea societies. By contrast, among the Machiguenga in Amazonian Peru, almost three quarters of the offers were a quarter of the pie or less and yet of 70 offers, there was just a single rejection, a pattern strikingly different from the student experiments conducted thus far. However, even among the Machiguenga, the mean offer was 27.5 percent, far more than would have maximized the proposer’s payoffs given the scant likelihood of a rejection.

Analysis of the experiments led us to the following conclusions: behaviors are highly variable across groups, not a single group approximated the behaviors implied by the self-interest axiom, and between-group differences in behavior seem to reflect differences in the kinds of social interaction experienced in the everyday life of the social group in question.

The evidence for economic conditions affecting behavioral norms is quite compelling. For example, the Aché in Paraguay share equally among all group members some kinds of food (meat and honey) acquired through hunting and gathering. Most Aché proposers contributed half of the pie or more. Similarly, among the Lamalera whale hunters of Indonesia, who hunt in large crews and divide their catch according to strict sharing rules, the proposer’s average allocation to the respondent was 58% of the pie. Moreover the Indonesian whale hunters played the game very differently from the Indonesian university students who were the subjects in another set of experiments (Cameron 1999). Indeed, where voluntary public goods provision was customary in real life (for example, the Harambee system, among the Orma, in Kenya), contributions in the experimental public goods game were patterned after actual contributions in the actual Harambee system. Those with more wealth contribute more. The wealthy behaved like the other Orma in the ultimatum game, for which there apparently was no everyday life analogue.

It seems likely that the correspondence between the typical livelihood of a group and its customary forms of interaction on the one hand, and the experimental behaviors of its subjects on the other, results from the fact that appropriate behavior is influenced by both custom and livelihood, and
these behaviors are then generalized and applied to novel situations such as our experiments.

Evidence that institutions serve as cues for appropriate behaviors comes from ultimatum game experiments with U.S. subjects in which simply naming the game “The Exchange Game,” or assigning the role of proposer to those who did well on a current affairs test, resulted in lower offers and a significant reduction in rejections of low offers (Hoffman et al. 1994b). If individuals cared only about their money payoffs, neither manipulation would have changed the game. The fact that significantly less strong reciprocity occurred in the “exchange game” and the current events test version suggests that social structure affects behavior in ways other than those captured by the money payoffs of the game, in this case by suggesting appropriate behavior (the “exchange game”) or identifying some individuals as “deserving” (the test manipulation).

Finally, experimental play in the ultimatum game and public goods with punishment game also suggest that institutions may influence behaviors in ways that go beyond the incentives and constraints that they implement. Recall that experimental subjects in the public goods with punishment game readily punish low contributors, despite the fact that in doing so they are adopting a strategy that is dominated in the game’s payoffs, meaning that there is some other strategy that would guarantee higher payoffs to the player irrespective what the other players do. Yet the same willingness to adopt a dominated strategy in the interests of the public good is uncommon in the standard public goods game, where there is no punishment option. Our interpretation is that, like institutions in natural settings, by providing the option of reducing the payoffs of others once one knows how much they contributed (the word “punishment” is never used) the game structure conveys information about appropriate behavior and influences beliefs about the actions of others.

Similarly, in the ultimatum game, people in the role of proposer often make an offer that approximately maximizes expected income from the game, the expectation being based on the ex post empirically observed rejection behavior of the respondents. But, in the role of respondent, the very same people rarely maximize expected income, for doing so would entail accepting any positive offer. For example, among the Hadza, hunter-gatherers whom we studied in Tanzania, the mean ultimatum game offer was almost exactly that which maximized expected income, but a quarter of all offers were rejected, and over two fifths of offers of 20% or less were
rejected. In this case the social roles created by the game, proposer and responder, apparently cue different behavioral reactions. Among the Hadza, in the role of proposer, considerations of fair treatment are apparently not salient, while in the role of responder, they are.

The fifteen small-scale society study described here was replicated by using improved methods and a partially overlapping set of small-scale societies with considerably sharpened results (Henrich et al., 2001, 2004).

5.11 Behavior is Conditioned on Group Membership

In experimental and natural settings, people often behave differently towards others, depending on the linguistic, ethnic, and religious groups to which they belong. People choose to associate with others who are similar to themselves in some salient respect (Lazarsfeld and Merton 1954, Thibaut and Kelly 1959, Homans 1961). Among the salient characteristics on which this choice operates are racial and ethnic identification, and religion (Berscheid and Walster 1969, Cohen 1977, Kandel 1978, Tajfel et al. 1971, Obot 1988). Conversely, people often seek to avoid interactions with those who are different from themselves.

Those who condition their behavior on the group membership of the other may do this because group membership is thought to provide information about the other’s likely behavior. Or, group membership may matter because people value the well-being of, or prefer to interact with members of some groups more than others. In first case the actor’s beliefs are involved. In the second case, group-sensitive preferences are at work. Group-sensitive preferences may be other-regarding (valuing the well-being of members of one’s own group, for example) or self-regarding preferences (e.g. experiencing anxiety in culturally unfamiliar interactions).

Laboratory experiments (with student subjects) have confirmed the salience of group membership in many settings. In the minimal group experiments initiated by Henry Tajfel and his colleagues (Tajfel et al. 1971), experimental subjects were assigned to groups on the basis of some trivial distinction (commonly their preference for paintings by Paul Klee over those of Wassily Kandinsky). In-group favoring behavior was quite pronounced in these experiments. Later prisoner’s dilemma and common pool resource game experiments found higher levels of cooperation when the players are members of the same minimal group than when they are
not members of the same group (Kramer and Brewer 1984). However, a series of experiments by Toshio Yamagishi (2003) and his associates show that experimental subjects’ allocations favor in-group members not because of altruistic sentiments towards those who are similar to themselves, but because they expected reciprocation from in-groupers and not from out-groupers.

In contrast to the minimal group experiments favored by psychologists and sociologists, behavioral economists have used the trust game, described in §5.11, introduced by Berg et al. (1995), generally with experimental subjects drawn from real-world ethnic groups. Player A is awarded a sum of money and given the opportunity to transfer any amount of this to Player B, knowing that the experimenter will triple the amount transferred (if A gives \( x \), B receives \( 3x \)). Player B then has the opportunity to return some the augmented transfer to Player A. This ends the game. Player A is sometimes called the “truster” and player B the “trustee.”

If A cared only about payoffs, and assumed that B had the same self-regarding preferences, A would transfer nothing for A would correctly infer that whatever B received would be kept rather than returned. When the game is played anonymously Player A typically contributes a significant amount, and significant amounts are returned by player B.

A number of experimenters have implemented the trust game played between subjects who were aware of the ethnic, religious, or linguistic identity of their partner. Fershtman et al. (2002) implemented this game in Belgium, played between students at Flemish and Walloon universities. Both Flemish and Walloon Player A’s make lower offers to outsiders than insiders but do not discriminate in favor of their own kind if the alternative is a Player B with undisclosed identity. When the same experiment is run in Israel, ultra orthodox Jews in the role of Player A give more to other ultra-orthodox Jews than to secular partners, but do not discriminate against secular partners by comparison with anonymous partners.

Discrimination against outsiders or in favor of insiders is far form ubiquitous, however. In another study using Belgian subjects, Bouckaert and Dhaene (2004) found no evidence of either type of discrimination in a trust game played by small businessmen of Belgian and Turkish origin. In the experiment just mentioned Yamagishi found that once subjects realized that they could not increase their own payoffs by contributing to their own group, they contributed equally to both groups. Other studies suggest that in some circumstances, ingroup favoritism is quite limited or even absent.
In a series of experiments in ethnically and linguistically divided poor neighborhoods of Kampala, Uganda, there was no evidence of favoritism towards co-ethnics in allocating sums or in choosing partners in a problem solving task. Remarkably, ethnic favoritism, which was absent in an anonymous dictator Game, was evident among some subjects when the dictator’s identity was public information. James Habyarimana and his coauthors (2009) explained that subjects believed that others would think less of them were they not to favor co-ethnics. A series of experiments with subjects drawn from differing ethno-linguistic groups in the highlands of Papua New Guinea (Bernhard et al. 2006, Efferson et al. 2010) provide further evidence on the relevance of group boundaries. We describe these in Chapter 9, as they have a direct bearing on the evolution of what we term parochial altruism modeled there.

Nonetheless, taking account of ethnic, racial and other characteristics of those with whom one interacts appears to be a common human trait. We seem quite attuned to noticing and treating as salient the ascriptive markers of group difference. For example, Americans of European and African origin are better at recognizing faces of their own ancestral group, and faces of their own group induce greater activation in the part of the brain associated with face recognition. Phelps et al. (2000) used brain imaging techniques (functional magnetic resonance imaging, fMRI) to study the neural substrates involved in the unconscious evaluation of Black and White social groups. They found that upon exposure to the (unfamiliar) faces of African American males (by comparison to the faces of European Americans), European American subjects exhibited heightened activation of the amygdala, an area of the brain associated with fear processing. Moreover, the extent of amygdala activation was correlated with an indirect (unconscious) measure of racial prejudice (the Implicit Association Test) but not with a direct (conscious expression) of race attitudes. Importantly, these patterns were not obtained when the stimulus faces belonged to familiar and positively regarded individuals (Colin Powell, Martin Luther King, Jr., Denzel Washington, eg.). Elizabeth Phelps and her coauthors see the

amygdala activation [as] reflections of social learning within a specific culture at a particular moment in the history of relations between social groups, [the effects of] cultural evaluations of social groups, personal experience with social group members, and one’s own group membership. (p. 734)
Due to the acute attention humans give to group boundaries, we might also be called “the parochial species.” The forms taken by parochialism today, religious intolerance, racism, xenophobia, vary across cultures and have evolved over time. But the various forms taken may share a common provenance, in the evolutionary processes that have made group boundaries salient to people. Like altruism, discriminatory preferences are an evolutionary puzzle, as they often impel people to forgo opportunities for beneficial exchanges and other interactions. We will address this puzzle in Chapter 9.

5.12 People Enjoy Cooperating and Punishing Free-riders

The most parsimonious and compelling explanation of behavior in the ultimatum game, public goods game, and other social dilemma experiments is that people think that cooperating is the right thing to do and enjoy doing it, and that they dislike unfair treatment and enjoy punishing those who violate norms of fairness. Some studies of collective action in natural settings are consistent with this view. An ethnographic study of people who exposed themselves to mortal risks in support of an agrarian insurgency against an authoritarian regime in El Salvador, for example, identified both ethical and religious commitments and the pleasure of seeking to rectify past injustices as a key motivation (Wood 2003).

Recent studies of brain functioning provide some support for this hedonic view of cooperative behavior. Using positron emission tomography (PET), functional magnetic resonance imaging (fMRI) and other techniques, neuroscientists, economists and others have begun to study the activation of the different brain areas of subjects playing experimental games (Fehr and Kosfeld 2005). There is some evidence, for example, that ultimatum game respondents who reject a low offer exhibit heightened activation of the bilateral anterior insula, an area associated with negative emotional states such as anger and disgust (Sanfey et al. 2003). Camerer et al. (2005) comment: “It is irresistible to speculate that the insula is a neural locus of the distaste for inequality and unfair treatment…”

Our view that subjects enjoy cooperation is consistent with the results of a series of experiments in which mutual cooperation is associated with elevated activity in one of the reward-related areas of the brain. Rilling et
al. (2004) found that mutual cooperation along with a monetary payoff enhances activity in the striatum, the brain area involved in processing rewards resulting from a decision, more than the same payoff resulting from performance of an individual task. Moreover, mutual cooperation with a human partner produces a higher level of striatum activation than does cooperation with a computer partner. deQuervain et al. (2004) studied brain activation of subjects in a social dilemma who had the opportunity to punish a partner who had abused their trust. Among those punishing trust violators, they found enhanced activity in the dorsal striatum. Moreover, those who inflicted more punishment exhibited higher levels of activation than did those punishing less. A related study by Singer found that male subjects (but not female) experienced pleasure, evidenced by activation in a reward processing part of the brain, the nucleus accumbens, rather than empathy, while observing pain inflicted on a partner who had defected in response to a cooperative offer by the subject in a sequential prisoner’s dilemma (Singer 2005).

The above studies do not suggest that cooperating and punishing defectors is innate. Some foods that evoke disgust in one culture are delicacies in others. Cross-cultural experimental evidence is consistent with the view that behaviors in social interactions that trigger aversive reactions likewise vary from one society to another. Our inference from these studies concerns how best to explain behavior, not whether it is genetically or culturally transmitted.

The field of neuroeconomics is still in its infancy and our understanding may be substantially modified by subsequent work. But, the evidence available to date suggests that the brain processes the punishment of defectors and the achievement of mutual cooperation much as it processes other pleasurable behaviors. If this view is correct, altruistic cooperation and the altruistic punishment of defectors need not be explained by constraints on behavior but rather by their status as objectives, pursued by reward-seeking individuals and thus an aspect of individual preferences. This does not mean that ethical values are unimportant. Quite the contrary, the experimental evidence that norm violators are punished supports Trivers’ (1971) notion that behavior is often motivated by moralistic aggression, an interpretation strengthened by the fact that altruistic punishment is directed not only toward those who harmed the punisher but also toward those who have harmed others.
5.13 Competing Explanations

The fact that subjects who had never contributed to a charity gave some of their endowment to an anonymous other in the dictator game is a reminder that one can never extrapolate directly from the laboratory to behavior in natural settings. But much stronger doubts have been raised against the experimental evidence. Some have suggested that the importance of social preferences is vastly overstated and distorted by questionable experimental design, especially not properly incentivizing subjects and not giving them sufficient time to learn the proper money-maximizing behavior. In their prominent critique of the experimental findings presented in this chapter, Ken Binmore and Avner Shaked maintain that the self-regarding Nash equilibrium is attained in “most games with money payoffs that have a unique Nash equilibrium, provided that the payoffs are sufficiently large and the subjects have ample time for trial-and-error learning. In spite of much rhetoric to the contrary, the one-shot prisoner’s dilemma is a case in point.” (Binmore and Shaked 2010). But, the experimental evidence suggests otherwise.

In the case of the one-shot ultimatum game (Slonim and Roth 1998) and the gift exchange game (Fehr and Touagareva 1995), other-regarding behavior remains even at high stakes and for experienced players, and there is to our knowledge no experimental evidence that indicates otherwise (other experiments agreeing with the above are Roth et al. (1991), Hoffman et al. (1994a), Straub and Murnighan (1995), and Cameron (1999)). The high-stakes version of the prisoner’s dilemma with a long learning period mentioned by Binmore and Shaked has not been experimentally implemented, to the best of our knowledge. The supporting evidence Binmore and Shaked cite, Ledyard (1995) and Sally (1995), in fact provide no evidence for their assertions. Ledyard (1995) reports that cooperation unravels in the repeated public goods game when punishment is not allowed. This is true, but we already have seen that this not due to players “learning” the Nash equilibrium, but rather to cooperators’ reactions to the free-riding of other subjects. Moreover, the study of the sequential prisoner’s dilemma described in §2.2 illustrates that a large fraction of subjects strictly prefer cooperating to defecting in the one-shot prisoner’s dilemma, and especially strongly so when they are guaranteed that their partner will cooperate (Kiyonari et al. 2000). In addition, Kollock (1997) allows subjects to rank possible outcomes in the game, and shows that subjects strongly prefer the outcome of mutual
cooperation. These results suggest that defection in the one-shot prisoner’s dilemma is often the result of aversion to betrayal rather than selfishness (Bohnet et al. 2008). In short, Binmore and Shaked’s generalization that subjects play the Nash equilibrium for self-regarding payoff-maximizers in simple games when this equilibrium is unique, the subjects are experienced, and the stakes are sufficiently high does not have empirical support.

It is worth noting that even were the degree of other-regarding behavior to decline with high stakes, as we have seen it did in the experiments of Andreoni and Miller (2002), this would not indicate the absence of social preferences, but rather that subjects are less likely to perform an action when it becomes more costly to do so. The fact that we would eat less ice cream at $10 a cone than at $3 a cone is not evidence against the hypothesis that we love ice cream. It indicates only that we have a limited budget and ice cream competes with other things we like for a share of our purchases.

Are there other plausible interpretations of cooperative behavior in the various games we have analyzed above? One might suggest that subjects simply do not understand the games, and cooperate by mistake. This is not very plausible, because the games are extremely simple and experimenters generally require subjects to exhibit understanding before permitting them to participate. Moreover, if failure to understand were the problem, subjects who play several ultimatum games in succession with different partners, for instance, should eventually learn to accept any positive offer, which they do not (Slonim and Roth 1998, List and Cherry 2000).

Another possibility is that because the anonymous, non-repeated interactions are not a part of everyday life, we should expect subjects to confuse the experimental environment with a repeated interaction, and for example, in the ultimatum game, to reject low offers in order to establish a reputation for hard bargaining.

But we do not believe that this argument is correct. Of course experimental subjects bring to the laboratory the moral sensibility and practical knowledge that their personal experiences and received cultural wisdom have conveyed. But, the confusion of a one-shot with a repeated interaction, or adhering to a maxim that says “act the same way in anonymous one-shots as in real-life repeated interactions” are not likely to be among these received wisdoms. We are very capable of distinguishing individuals with whom we are likely to have many future interactions, from those with whom future interactions are less likely. Indeed, the sharp end-game drop off in contributions exhibited in figures 3.2 and 5.2 shows that experimental
subjects are very sensitive to this distinction, cooperating much more if they expect frequent future interactions than if future interactions will not occur (Keser and van Winden 2000, Gächter and Falk 2002).

Other data support the notion that responders reject positive offers, not because they are confused or falsely believe they can establish a reputation for hard bargaining thereby, but simply because they want to punish an unfair proposer. For instance, in a variant of the game in which a rejection leads to the responder getting nothing, but allows the proposer to keep the share he suggested for himself, respondents rarely reject offers, and proposers make considerably smaller (but still positive) offers (Bolton and Zwick 1995).

A quite different critique of the experimental evidence is the ubiquity of framing effects, which are payoff irrelevant facts that affect behavior. For instance Andreoni (1995b) shows that, when a public goods game is framed as an “individual investment game” in which subjects can invest in their private accounts, even though this imposes a cost on the other players, subjects exhibit little cooperation. Because the actual payoffs to the game are the same whether the experimenter stresses the gains to social cooperation or the gains to personal investment, it appears that the preference for cooperation is a fragile commitment easily manipulated by the experimenter.

However, it would be very odd if other-regarding preferences, such as for fairness, reciprocity, or honesty, were not sensitive to framing affects. The reason is that the ethical values an individual applies to a given situation depend on how the individual conceptualizes the social norms appropriate in that situation, and this conceptualization is subject to experimenter manipulation. The fact, then, that the experimenter can find a set of conditions under which subjects behave selfishly is not an argument against the importance of other-regarding preferences. For instance, in Andreoni (1995b), the instructions to subjects naturally led each one to believe that the others would stress personal over collective gain, and hence the principles of strong reciprocity would lead each player to fulfill the experimenter’s prophesy. If the subjects had been permitted to play the game repeated with the same partners, they might have overcome this initial reaction, but experimental conditions precluded groups lasting more than one period.

A final critique is that other-regarding and moral preferences are confused expressions of a desire to maintain a good reputation in the eyes of others. We will call this the “reputation” model of social preferences. Note that the critique is not that establishing a reputation in the game may be a way of gaining higher payoffs in some later interactions, as ones’ behavior in
the game is anonymous. Rather, the reputation model’s contention is that people are very concerned about the way they are evaluated by others, independent of any material reward or punishment they might receive by being judged. This view has considerable empirical support. As we have already seen, Masclet et al. (2003) showed that considerable cooperation could be induced in a public goods game where punishment consisted merely of being assigned “negative points” for free-riding. Similarly, Rege and Telle (2004) showed that having participants in a public goods game write their contributions on an announcement board in full view of the other participants increased contributions markedly, even though the subject were unlikely ever to meet again.

Perhaps the most dramatic effect of this type is the tendency of subjects in perfectly anonymous settings to perceive the presence of “imaginary” viewers, as revealed by the so-called eye-spot experiments. For instance, Bateson et al. (2006) studied contributions to an “honor box” for the use of tea, coffee, and milk in the common kitchen of the 48 members of the Division of Psychology at the University of Newcastle over a period of ten weeks. In alternating weeks, they placed a picture of flowers and a picture of a person’s eyes looking straight at the viewer, above the note to contribute to the honor box. They found that on average, people paid 2.76 times as much under the eye-spot condition than under the flowers condition. The experimenters conclude that this evidence, which is supported by other researchers (Haley and Fessler 2005, Burnham and Hare 2007), suggests that “the self-interested motive of reputation maintenance may be sufficient to explain cooperation in the absence of direct return.” (p. 413)

However one interprets the eye-spot phenomenon in dictator game experiments, there is no indication that eye-spots have any efficacy in social situations with strategic interaction among participants. Indeed, in an eye-spot experiment using the trust game, the presence or absence of eye-spots has no statistical effect on the behavior of the subjects (Schneider and Fehr 2010). In addition, it is incorrect to infer from the fact that people act more generously when there appear to be witnesses that people exhibit other-regarding preferences only when they believe, consciously or otherwise, that they are being observed. The above evidence is completely consistent with our view that individuals have moral values that they uphold for their own sake, although their self-assessment as moral beings is highly sensitive to how they fare in the eyes of others.
The most likely source of this heightened sensitivity to the opinion of others was investigated thoroughly long ago by the sociologist Charles Horton Cooley (1902), who coined the term “looking-glass self” to describe the dynamic whereby individual psychic well-being depends on self-esteem, and self-esteem is the product of being esteemed by others with whom one shares a moral community. According to Cooley, and after him the renowned sociologist George Herbert Mead (1934), we strive to please others not only for reputation (material reward in either present or future), but also because our self-esteem depends on others’ evaluation of us.

There have been several interpretations of the looking-glass self as a model of self-esteem. The one best supported by the evidence (Leung and Martin 2003) is the internalization of norms model according to which we internalize norms that provide for us moral and prosocial preferences, and our self-esteem depends on meeting moral and prosocial expectations. While some individuals are capable of maintaining high self-esteem from personal self-assessment, most individuals are acutely dependent upon the positive evaluation of their behavior by others. The looking-glass self is thus an amalgam of personal self-assessment and the assessment of others.

It is entirely plausible, therefore, that even the subconscious neural cues of being watched in the eye-spot experiments would lead many individuals (those dependent upon the evaluation of others for self-esteem) to increase their other-regarding behavior. This, however, in no way supports the notion that experimental evidence for social preferences is spurious.

Many economists, biologists and others will assert, as they have for at least a century, that altruism beyond one’s immediate family members is highly exceptional and ephemeral. The experimental evidence of the last two decades tells strongly against this view. But the belief that self-interest is unrivaled among human motives has never depended on empirical tests, rather it has appeared self-evident because the evolution of the human species by a process of natural selection was bound to produce a selfish animal. In the absence of a plausible evolutionary explanation of the origin of altruistic preferences (excepting genealogical kin), the self-interest assumption was commonly accepted by default. But, as we will see in the next and subsequent chapters, the idea that selfish genes must produce selfish individuals is false.
Multi-level Selection and the Evolution of Altruism

Selfish and contentious people will not cohere, and without coherance, nothing can be effected. A tribe possessing... a greater number of courageous, sympathetic and faithful members, who were always ready to warn each other of danger, to aid and defend each other... would spread and be victorious over other tribes... Thus the social and moral qualities would tend slowly to advance and be diffused throughout the world.


In light of what we know about other animals and about human evolution, how could the preferences documented in the previous chapter have become common among humans? Both the experiments and the models surveyed in chapters 3 and 4 make it clear that neither “self interest with a long time horizon” nor a desire to care for family members can explain the distinctive nature of human cooperation, namely, that it extends far beyond the family and is often motivated by ethical commitments or concerns for others, that lead people to cooperate with others even when not doing so would raise their material payoffs. Given the tendency of people to copy the successful and of natural selection to favor the more fit, how did our altruistic preferences overcome the cultural and biological evolutionary handicaps entailed by the reduced payoffs that they entailed?

Our explanation hinges on three facts. First, group living is essential to human survival. Second, groups differ in their evolutionary success, some expanding and dividing many times, while other groups are absorbed into more successful groups or pass out of existence in warfare or during environmental crises. Third, groups in which altruistic and other social preferences are common tend to cooperate, and cooperative groups tend to prevail in intergroup competition, and to survive environmental crises.
Differential group success therefore plays a central role in the evolution of human behaviors and institutions, less successful groups copying the more successful or being eliminated by them. Examples of this process include the peopling many parts of the world by individuals of European ancestry and the associated spread of European customs and institutions in the past half millennium, and the spread of agriculture and its associated novel systems of social organization and behavior from the Middle East to Europe beginning eleven millennia ago.

The fact that more cooperative groups tended to survive and expand explains two key facts. The first is that altruistic individuals among our ancestors enjoyed enhanced reproductive success, resulting in the spread of altruism as a distinctive human trait. The second is that our altruistic dispositions motivate us to care about and help unrelated others. The likelihood that the group-structured nature of populations and group competition would strongly affect evolution has long been recognized not only by sociologists and historians (Tilly 1981, Parsons 1964) but also by biologists (Lewontin 1965,1970, Alexander 1979, Wilson 1977, Durham 1991, Dunbar 1993, Laland and Feldman 2004). But, until recently, most biologists have concluded that group-level effects that would favor the spread of genes contributing to altruistic behavior cannot offset the effects of individual within-group selection operating against altruists, except where special circumstances heighten and sustain genetic differences between groups relative to differences within the group (Williams 1966, Crow and Kimura 1970, Boorman and Levitt 1973, Maynard Smith 1976). The reason is that the speed of an evolutionary process is proportional to the differences on which it works, so in order for between-group selection to outrun within-group selection, between-group differences must be substantial. But, gene flow due to group exogamy and other reasons for migration are thought to preclude this.

However, beginning with Darwin, whose position is clear from the head quote of this chapter, a number of evolutionary thinkers have suggested that human evolution might provide an exception to this negative assessment of the force of group level, or as we shall call it, multi-level selection. William Hamilton (1975):331 summarized Darwin’s view as follows: “He saw that such traits [as]…courage and self-sacrifice…would naturally be counter-selected within a social group, whereas in competition between groups the groups with the most of such qualities would be the ones best fitted to survive and increase.”
In *The Causes of Evolution*, J. B. S. Haldane (1932) had already provided a plausible mechanism for how this might come about. He suggested that in population of small endogamous ‘tribes,’ an altruistic trait might evolve because the ‘tribe splitting’ that occurs when successful groups reach a certain size would by chance create a few successor groups with a very high frequency of altruists, reducing within-group differences and increasing between-group differences, a process very similar to that we will model in chapters 8 and 9. The modest size of typical human groups during most of our evolution thus could play a crucial role in the chance occurrence of one or more groups with a high fraction of altruists, which would then, as Darwin said, “spread and be victorious over other tribes.” Haldane concluded: “evolution in large random-mating populations...is not representative of evolution in general, and perhaps gives a false impression of the events occurring in less numerous species....Our ancestors were mostly rather rare creatures.” (p. 213-14) William Hamilton (1975) took up Haldane’s suggestion, adding that if the allocation of members to successor groups following ‘tribe splitting’ was not random but was rather what he called ‘associative,’ (p. 137) between-group differences would be even greater and multi-level selection pressures would be further enhanced.

More recent research also suggests that impediments to multi-level selection may be less general than was once thought (Uyenoyama and Feldman 1980, Harpending and Rogers 1987). A number of writers have pointed out that multi-level selection may be of considerably greater importance among most humans than among other animals given the advanced level of human cognitive and linguistic capabilities and consequent capacity to maintain group boundaries and to formulate general rules of behavior for large groups, and the resulting substantial influence of cultural inheritance on human behavior (Alexander 1987, Cavalli-Sforza and Feldman 1973, Boyd and Richerson 1985,1990, Sober and Wilson 1994, Boehm 1997).

Among the consequences of these distinctive human capacities are the suppression of within-group phenotypic differences through egalitarianism, coinsurance, consensus decision making and conformist cultural transmission. Insider biases and individual preferences to interact with like individuals lead to large between-group differences in behavior, and to a lesser but not negligible extend, in genetic predispositions, too, as we will see. These insider biases result in supporting high levels of assortative interactions both within and between groups and frequent between-group conflict. Other animals do some of these things, but none does all of them on a hu-
human scale. All of these aspects of human social life enhance the force of between-group selection relative to within-group selection.

Here we explain how multi-level selection works and its relationship to other biological models of the evolution of human behavior thus far introduced in Chapter 3. Group selection may take both a strong form, in which group competition offsets the selective pressures operating against altruistic members within a group, and a weak form, in which group competition selects among stable within-group equilibria, meaning that within-group selection pressures against altruistic members are absent, as a result favoring groups with a higher level of cooperation. In §6.3 we show that this group selection model shares an important characteristic with the other biologically inspired models reviewed in Chapter 3, kin altruism and reciprocal altruism: they all work because altruists are more likely than non-altruists to interact with individuals who help them. The models differ in the processes that account for this behavioral assortment benefiting the altruistic types. In part for this reason the models also differ in the proximate motives likely to account for helping behavior, providing further clues, which we consider in the concluding section, to the puzzle of how human cooperation evolved.

### 6.1 Multi-level Selection

The multi-level selection model works because members of predominantly altruistic groups have above average fitness and thus contribute disproportionately to the next generation. Here we apply multi-level selection to a process of genetic transmission, but the same model applies to any process of selection based on the differential replication of traits over time. To see how group selection works, consider a single altruistic behavior that is the expression of an ‘altruistic allele,’ the quotation marks being a reminder that altruistic behaviors are unlikely to be the expression of a single allele.

Thus, we consider a trait that may (A) or may not (N) be present in each individual in a large population that is subdivided into a number of reproductively somewhat isolated groups, commonly know as demes. Suppose altruistic behavior costs the individual c and confers a total benefit of b on a randomly selected member of the group, where c and b are both measured in units of fitness, and \( b > c > 0 \). It follows that a member in a group composed entirely of A’s has a payoff of the amount \( b - c > 0 \), greater than a member of a group composed of all N’s. In any mixed group, however,
the expected payoff to $A$’s will be lower than that of the $N$’s by an amount $c$. These payoffs are the same as in Figure 3.1.

Let $p_{ij} = 1$ indicate that individual $i$ in group $j$ is an $A$, with $p_{ij} = 0$ if $i$ is an $N$. Let $p$ represent the fraction of $A$’s in the population at the start of a given time period, and let $p'$ be the fraction of $A$’s at the start of the next period. Define $w_{ij}$ as the expected fitness of an individual of type $i$ in group $j$ and let

$$w_{ij} = \beta_o + p_j \beta_g + p_{ij} \beta_i$$  \hspace{1cm} (6.1)

where $\beta_g$ and $\beta_i$ are the effects on $w_{ij}$ of $p_j$ (the frequency of $A$ in the group and $p_{ij}$) the presence of the $A$ allele in individual $i$, respectively, and $\beta_o$ is a baseline replication rate based on factors not considered here. In equation 6.1 the two effects, between- and within-group selection, are separable, so that the size of the group effect on an individual is independent of whether the individual is altruistic or not, and the individual effect of being an altruist is independent of the frequency of altruists in the group. The assumption that the two effects are separable is unrealistic, but it allows an illuminating simplification in this case.

George Price (1972) showed that the change in the frequency of altruists in the population, $\Delta p \equiv p' - p$, can be partitioned into between-group and within-group effects. We prove this in Appendix A9, where we show that the partition can be written as

$$w \Delta p = \text{var}(p_j) \beta_G + \beta_i \overline{\text{var}}(p_{ij}).$$  \hspace{1cm} (6.2)

where $w$ is the population-wide average of the number of replicas, which we normalize to unity, as we assume the population size is constant, and $\overline{\text{var}}(p_{ij})$ is given by

$$\overline{\text{var}}(p_{ij}) = \sum_j f_j \text{var}(p_{ij}),$$  \hspace{1cm} (6.3)

where $f_j$ is the fraction of the population in group $j$. The terms $\text{var}(p_j)$ and $\overline{\text{var}}(p_{ij})$ respectively are the between-deme genetic variance and weighted average within-deme genetic variance (the weights based on group size). The coefficient $\beta_G$ is the effect of variation in $p_j$ on the average fitness of members of deme $j$ (that is, on $w_j$), which is determined as follows. If we sum 6.1 over the individuals in a group and divide by group size, we see that

$$w_j = \beta_o + p_j (\beta_g + \beta_i).$$  \hspace{1cm} (6.4)
So \( \beta_G \equiv dw_j/dp_j = \beta_i + \beta_g \).

Recall that a behavior is altruistic if adopting it, hypothetically switching from an N to an A, lowers one’s expected fitness while increasing the average fitness of one’s group. Our definition of altruism thus maps neatly onto the terms of the Price equation, and we are interested in the cases where \( \beta_G > 0 \) (altruism is group-beneficial) and

\[
\beta_i = \frac{dw_{ij}}{dp_{ij}} = \frac{\partial w_{ij}}{\partial p_{ij}} + \frac{dp_j}{dp_{ij}} g = -c + b/n < 0,
\]

where the first term is the effect of an individual switching from N to A while holding constant the fraction the group that are A’s (hypothetically accomplished by some other individual switching in the other direction) while the second term on the right hand side captures the effect on the individual’s fitness of the change in the fraction of the group that are A’s independently of his trait, namely that he is the beneficiary of his own altruistic act with probability \( dp_j/dp_{ij} = 1/n \) where \( n \) is deme size. To pose the problem of the evolution of altruism in the starkest and simplest manner, we here assume that \( n \) is sufficiently large that the second term partially offsetting the direct costs of the altruistic act may be ignored. In our empirical applications of this model we take account of the likely modest size of prehistoric demes.

The separability assumption now allows us to represent equation 6.2 in terms of the payoffs: \( \beta_i = -c \) and \( \beta_g = b \). Thus equation 6.2 can be written

\[
w \Delta p = \text{var}(p_j)(b - c) - c \sum_i \text{var}(p_{ij}). \tag{6.5}
\]

The first term captures the group effect, which is positive, by the definition of altruism, while the second represents the effect of within-group selection, which is negative, also by the definition of altruism. Setting aside degenerate cases such as zero variances, it follows that the frequency of the trait will be stationary (i.e., \( \Delta p = 0 \)) where these two terms in equation 6.4 are of equal absolute magnitude (assuming that the \( \beta \)'s and variances making up these terms are themselves stationary). Because the second term is negative, the frequency of the A-trait within all groups will fall over time. But as \( b - c \) is positive, this tendency will be offset by the decline in the size of groups with low frequencies of the trait and the expansion of groups with many altruists.
Let us define the variance ratio $F_{ST}$ as the ratio of the between-group variance in the fraction of altruists to the total population variance, which is the within-group plus the between-group variance of the fraction of altruists, or

$$F_{ST} = \frac{\text{var}(p_j)}{\text{var}(p_{ij}) + \text{var}(p_j)} \quad (6.6)$$

The variance ratio is thus a population-wide measure of the degree of non-randomness in who interacts with whom, resulting from the tendency of altruists to find themselves disproportionately in groups with many other altruists. The variance ratio $F_{ST}$ is Sewall Wright’s inbreeding coefficient (Wright 1922), measuring the degree of genetic differentiation among groups.

This ratio measures the difference between the probability of being paired with an altruist conditional on being an altruist, $P(A|A)$, and conditional on being a non-altruist, $P(A|N)$, that arises because the population is group-structured (Crow and Kimura 1970). Thus,

$$F_{ST} = P(A|A) - P(A|N). \quad (6.7)$$

An example illustrating this result is given in §A9. Being “paired with an altruist” means being the recipient of the randomly assigned benefit, $b$, that altruists confer on fellow group members. This is more likely to occur if there are many A’s in one’s group. If A’s tend to be in groups with many other A’s, they enjoy an advantage over N’s. $F_{ST}$ measures the extent of this advantage. Rearranging terms in (6.2), we see that the condition for $\Delta p$ to be zero so that the fraction of altruists in the population is stationary, is

$$F_{ST} = \frac{\beta_i}{\beta_g + \beta_i}, \quad (6.8)$$

or

$$F_{ST} = \frac{c}{b} \quad (6.9)$$

If $F_{ST} > c/b$, the fraction of altruists may be expected to increase and if $F_{ST} < c/b$, it decreases.

Equation 6.9 also indicates the most costly form of altruism that may proliferate by this method. When the variance among group means is zero (so that $F_{ST} = 0$), A’s no longer have the advantage of being in groups with disproportionately many A’s. In this case multi-level selection is inoperative, so only a costless form of group beneficial behavior could proliferate.
By contrast when $\text{var}(p_{ij}) = 0$ for all $j$, groups are either all A or all N, and one meets only one’s own type, independently of the composition of the total population. In this case, within-group selection is absent and between-group selection is the only selective force at work. Thus, the force of multi-level selection will depend on the magnitude of the group benefit relative to the individual cost ($b$ and $c$) and the degree to which groups differ in their frequency of the trait, relative to the within-group variance of the trait.

Figure 6.1 shows how the group structure of the population may overcome the disadvantage of bearing the costs of altruistic behaviors and indicates the variance ratio $F_{ST}$, namely the difference $P(A|A) - P(A|N)$, that is just sufficient to equate the expected fitness of the two types and thus to maintain a stationary value of $p$. As is evident from the figure, this is the $F_{ST}$ that satisfies equation 6.9, namely $Fb = c$. The slope of both expected fitness functions is $b$ and the distance between them is $c$. How large $P(A|A) - P(A|N)$ must be depends, as we have seen and as the figure makes clear, on the costs and benefits of the altruistic behavior, $b$ and $c$. It is clear from the figure that for a given level of $b$ (slope of the lines), the greater is the cost of altruism, $c$ (the vertical distance between the lines) the greater must be the degree of positive assortment $P(A|A) - P(A|N)$ in order for altruism to proliferate in the population. In Appendix A9 we provide an numerical example that will clarify how the model works.

The Price equation does not represent a complete dynamical system giving the movements of $p$. This would require a set of equations giving the movements over time of the between- and within-group variances. It is easy to check, for example, that the variances given in the numerical example in the Appendix that render $p$ stationary are not themselves stationary. Because they will be different the next period, the values of $b$ and $c$ that made $p$ stationary in the period under study will not ensure stationarity in subsequent periods. Except in degenerate cases of little interest, equations giving the movement of the relevant variances over time are not attainable. This is the reason that exploring evolutionary processes under the influence of group selection typically requires that we resort (in subsequent chapters) to agent-based models.
Figure 6.1. **The Evolution of an Altruistic Trait in a Group-structured Population.** The vertical axis measures expected fitness minus the baseline fitness $\beta_o$. If the population structure’s variance ratio is such that the difference in the conditional probabilities of being paired with an A, $P(A|A) - P(A|N)$, is as shown, $p$ is stationary, because the expected fitness of the two types, $w_j^A = bP(A|A) - c - \beta_o$ and $w_j^N = bP(A|N) - \beta_o$ are equal.

### 6.2 Equilibrium Selection

The above model illustrates what we call strong multi-level selection, which occurs when the second term in the Price equation is negative, indicating that those with group-beneficial traits are altruistic and hence would attain higher fitness by abandoning their helping behaviors. But the second term may be zero, so that those who confer benefits on others suffer no disadvantage within groups while profiting from the group beneficial effects of their behaviors. When this is the case, we say weak multi-level selection is operative. The cooperative traits that may be supported in the equilibria selected by this process are not altruistic, because cooperators would not gain in fitness terms were they to defect. There are two main cases of weak multi-level selection.

The first type of weak multi-level selection occurs when both the A-trait and the N-trait are evolutionarily stable strategies, so that two stable within-group equilibria exist, one with all $N$’s and the other with all $A$’s, each of which cannot be invaded by a small group of migrants or mutants exhibiting the other type of behavior. This was the case, as we have seen, when the conditions for the proliferation of cooperation by means of reciprocal altruism, indirect reciprocity or costly signaling obtain. These models provide mechanisms by which a (non-adaptive) cooperative behavior could be stabilized within a population. Note that in this case the second term of the Price equa-
tion is zero because if $p_j = 1$ or $p_j = w$, then $\var(p_{ij}) = p_j(1 - p_j) = 0$. This would also be the case, for example in the repeated game setting in which both unconditional defection and conditional cooperation are mutual best responses. But here, we model the case of weak group selection by extending the above model of a one-shot interaction between $A$’s and $N$’s. Two stable equilibria could exist in this case, for example if, following an $A$-$N$ interaction, the $A$’s in the group collectively attempted to punish the $N$. The $A$’s would then be akin to the strong reciprocators whose preferences were described in the previous chapter. Suppose their attempt to punish the $N$ succeeds with a likelihood equal to the fraction of $A$’s in the population, and if successful the cost imposed on the $N$ is $c_p$. Assume each $A$ incurs a cost of $k$ in attempting to punish the $N$. Then, the expected fitness of the $A$’s in group $j$ is

$$w_j^A = p_j b - c - k(1 - p_j) + \beta_o,$$

giving the expected fitness functions within a single group illustrated in upper panel of Figure 6.2, for the case where $c_p > c$, so the target of the punishment bears a greater cost than does an individual punisher. Note that if $p > p^*$, $A$’s will have higher fitness and thus eventually eliminate the $N$’s, while if the reverse inequality holds, $N$’s will have higher fitness and thus eventually eliminate the $A$’s. A population composed of many groups with this interaction structure could contain some with all $A$’s and some with all $N$’s. Because groups are homogeneous, they would remain so, but the all $A$ groups would have higher average fitness, and if the total population is constant, the $N$’s would eventually be eliminated. Boyd and Richerson (1985) showed that conformist cultural transmission, a tendency to copy the more common behaviors, can have the same effect, giving rise to homogeneous groups among which weak group selection can support the proliferation of an altruistic trait.

In the second type of weak multi-level selection, cooperative individuals coexist with others in a stable within-group equilibrium. Multi-level selection works because groups in which the equilibrium number of cooperators is greater prevail in competition with other groups. In this case, at the equilibrium for a particular group, the second term of the right hand side of the
price equation is zero because the fitness of the two types must be equal (otherwise it is not an equilibrium) and as a result, the $A$ trait suffers no fitness deficit. An example is in a model of cultural rather than genetic evolution in which non-altruists may turn into altruists, by being socialized through group rituals to behave altruistically, and altruists can revert back to non-altruists, attracted by the possibility of not paying the cooperation cost $c$. Equilibrium now occurs when the two movements from altruist to non-altruist and back lead to equal fitness of both phenotypes. This is shown in the lower panel of Figure 6.2, which depicts the model just described. The equations on which the bottom panel is based are in Appendix A10. We present a more complete model along these lines in Chapter 11.

Figure 6.2. **Weak Multi-level Selection.** The horizontal axis exhibits the fraction of atheists in the group. In the upper panel $p_j^*$ is an unstable equilibrium and constitutes the boundary between the basin of attraction of the all-$N$ and the all-$A$ equilibria. In the lower panel, $p_j^*$ is a stable equilibrium. In both cases within-group selection will result in a stationary distribution of $A$’s and $N$’s with no within-group selection operating against the $A$’s. By contrast, Figure 6.1 illustrates strong multi-level selection.

In the upper panel, depicting the first model in which both cooperation and non-cooperation are evolutionarily stable strategies, $F_{ST} = 1$, reflecting
the fact that all of the variance is between groups and none is within, so the second term on the right hand side of Price equation is zero. This, or equivalently condition 6.9, tells us that altruism can proliferate no matter how small is the excess of benefits to others over cost to self. In the lower panel in equilibrium (that is, at $p_j^*$) $\beta_i = 0$, as there is no within-group selection operating against the $A$’s, the second right hand side term in the Price equation is zero in this case too and equation 6.7 tells us that any positive $F_{ST}$ will be sufficient to promote the spread of altruism.

### 6.3 Positive Assortment

In the multi-selection model, as in the kin altruism models, the evolution of helping behaviors requires that those with a predisposition to help others receive help from those with whom they interact more frequently than would occur by chance. How this positive assortment comes about differs from case to case. Eshel and Cavalli-Sforza (1982) provide a glimpse of the variety of causes of assortment: “Kin, deme, niche, and social group structure, neighborhood effect, idiosyncratic behavior, and discrimination in the choice of companions are some of the possible sources of deviation from randomness…critical for understanding the evolutionary stability of social structure.”

For the group selection model, we have equation 6.9, which can be written:

$$P(A|A) - P(A|N) > \frac{c}{b}.$$

This is a generalization of Hamilton’s rule for the degree of positive assortment permitting an altruistic trait to proliferate when rare. This can be seen by returning to Figure 6.1 and noting that the same model applies to kin selection using the fact that the degree of genetic relatedness between the two members of an interacting pair $P(A|A) - P(A|N)$, which is equal to $r$. $P(A|A)$ is the probability that the bearer of an altruistic allele will interact with another bearer of that allele, while $P(A|N)$ is the probability that the bearer of the non-altruistic allele is paired with a bearer of the altruistic allele. The former will exceed the latter, reflecting positive assortment, if interactions are among individuals of common recent genealogical descent. Because the processes contributing to positive assortment among close descendant kin on the one hand and in group-structured populations on the other differ in important ways, we refer to evolutionary process based on
the former as kin selection, and use the relatedness term $r$, while we refer to the latter as multi-level selection, and use the variance ratio $F_{ST}$.

Surprisingly, the same expression gives the condition for reciprocal altruism to evolve even when the process of dyadic pairing to interact is random. This has been shown in a number of ways, beginning with Queller (1985) and including Nee (1989) and Fletcher and Zwick (2006). Here, the altruists are the Nice Conditional Cooperators, while the non-altruists unconditionally defect. In this interaction, A’s and N’s will be paired with an A with the same probability, $p$. But positive assortment of the A allele and the cooperative phenotype occurs nonetheless because an N paired with an A will benefit from the partner’s cooperation for just one period because the A, a Conditional Cooperator, will switch to defect after the first round. By contrast, the A paired with an A will benefit from the partner’s cooperation for as many periods as the interaction endures. The key to the evolutionary success of the Conditional Cooperator in this situation is, as Fletcher and Zwick explain, “there must be sufficient positive assortment between individuals with the altruistic genotype… and the helping phenotypes of others they interact with.” (Fletcher and Zwick 2006):253.

Thus, suppose we observe a large randomly-paired population in which the fraction of Conditional Cooperators is $p$, and we count the periods in which an A and an N respectively benefit from the cooperation of a partner expressed as a fraction of the total periods of interaction for each type. The former will exceed $p$ while the latter will be less than $p$. The longer the interaction endures, the greater will be the degree of positive assortment between altruistic genotypes and helping phenotypes measured by this difference.

It remains to show that the condition for Nice Conditional Cooperator to be an evolutionary stable strategy is also just another version of $P(A|A) - P(A|N) > c/b$. Recall that for the case of a repeated prisoner’s dilemma, this condition (equation 3.2) is $\delta > c/b$ where $\delta$ is the probability of continuation of the interaction at the end of each interaction, and $b$ and $c$ are respectively the benefits and costs of the cooperative act undertaken by the reciprocal altruist (A). Suppose that the expected total duration of an interaction from initiation to termination is normalized to one and that opportunities to alter one’s action occur at the end of every period, each of which is a fraction, $1 - \delta$, of the expected duration of the interaction. Thus there are $1/(1 - \delta)$ periods in the expected duration, and the probability of continuation at the end of each is $\delta$. We measure the advantageous pairing
of the A types by the fraction of the expected total duration of the interaction in which they experience helping from the individual with whom they are paired. To determine if A (all Nice Conditionally Cooperators) is an ESS, we study the difference in the number of periods in which an A and an N respectively will experience cooperation from its partner. The duration of cooperation enjoyed by an A is $p$, because with probability $p$ the A is paired with an A in which case they both cooperate for the expected duration of the interaction, which is 1. The corresponding duration of cooperation enjoyed by an N is $p(1 - \delta)$ because with probability $p$ the N will be paired with an A, who cooperates during the first period, which is a fraction $(1 - \delta)$ of the expected duration, and then defects for all subsequent periods. Setting $p = 1$ to study the evolutionary stability of the all Conditionally Cooperator equilibrium and letting $P(C|A)$ and $P(C|N)$ represent the expected duration in which an A and an N respectively experience cooperation from its partner, we have: $P(C|A) - P(C|N) = \delta$. Thus $\delta$ is a measure of positive assortment given by the difference in the frequency of cooperative actions by one’s partner, conditional on one’s type. So we can rewrite the condition for Conditionally Cooperate to be an ESS as $P(C|A) - P(C|N) > c/b$.

This demonstrates that the condition for helping behaviors to evolve by repeated interactions is the same as the condition for altruistic cooperation to evolve by means of kin selection or group selection. An ingenious set of papers (Ohtsuki et al. 2006) suggests yet another example of this logic, and one that captures essential aspects of human society. Until now we have studied just two types of social structure: families, meaning kin of recent common descent, and demes, that is sub-populations within a larger population. But, within any group, some individuals are closer to some than to others, and this social proximity should make a difference in evolutionary dynamics. Ohtsuki and coauthors show that it does.

Suppose Cooperators and Defectors are arrayed on a network and each plays a public goods game with all of the $k$ other “neighbors” with whom they are directly linked. As above, at a cost of $c$, each cooperative individual contributes an amount $b$ to be shared among its $k$ neighbors, whether they are Cooperators (C) or Defectors (D). Defectors bear no costs and contribute no benefits. Individuals periodically may alter their strategy by adopting a strategy of a neighbor, with a probability proportional to the relative fitness of that neighbor. Thus, the individual will adopt C with probability equal to the total payoffs of neighboring C’s divided by the total payoffs of all neighbors.
Through extensive simulations using a wide range of network structures, Ohtsuki and his coauthors find that the Cooperators proliferate if \(1/k > c/b\) where \(k\) is the average number of neighbors (the degree) of the players. This surprising result occurs because, for reasons that are not fully understood as yet, by chance there will be one more C in the neighborhood of C’s than in the neighborhood of D’s independently of the size of the neighborhood. Thus, the smaller the neighborhood, the more different in their compositions are they. It seems likely that the fact that smaller neighborhoods are (relatively) more different one from another than are larger ones arises by a process similar to sampling error where differences in the means of small samples exceed differences in the means of larger samples. Whatever the explanation of this regularity, the probability that any given one of my \(k\) neighbors will be a C differs by \(1/k\) depending on whether I am a C or a D. Thus, \(1/k\) is exactly equal to \(P(A|A) - P(A|N)\), affirming the positive assortment logic already demonstrated in the other models.

Whether this model explains observed forms of human cooperation is open to some question, however. The networks of foraging and other small scale societies are typically quite large often including virtually all adult members of a group. If average degree (\(k\)) of an individual were 20, altruism could not proliferate by this mechanism unless costs did not exceed five percent of the benefits. The average number of exchange partners in the Southern African !Kung coinsurance (so called \(hxaro\) networks described in §7.1 is 17, while the families among whom food is shared on a regular basis among both the Paraguayan Aché exceeds ten even for small packages, and 20 for large packages (Wiessner 2002, Kaplan and Hill 1985a). Average degree in the coinsurance networks among the Pokot and Himba herders are much larger than this (Bollig 2006). Thus, in order for a form of altruistic cooperation to be spread by this mechanism, it would have to have truly extraordinary ratios of benefit to cost.

Not surprisingly, positive assortment is also at work in the indirect reciprocity and costly signaling models. In all cases the degree of positive assortment determines the critical cost benefit ratio for the proliferation of altruism. The differences among the models arise in the way that positive assortment comes about. These results are summarized in Table 6.1.
Each model explains why humans might have come to behave cooperatively while differing in the sociological, strategic and demographic mechanisms accounting for positive assortment and remaining silent as to the proximate motives leading humans to engage in helping, and sometimes altruistic, behaviors. We have seen that positive assortment is a common feature of all of these models, so the distinctive character of each must lie in the processes
explaining how positive assortment comes about, or in the kinds of social preferences likely to result.

Thus, for example, if kin altruism were the main reason for the evolutionary success of helping behaviors, the motives involved in helping would include a love for one’s children and other close genetic relatives and a concern for their well-being sufficient to motivate self-sacrifice. Similarly, we would expect that if the sole mechanisms contributing to the evolution of helping behavior were reciprocal altruism, indirect reciprocity or costly signaling, the proximate motives for helping and the cognitive processes activating them would involve individual advantage, accompanied by detailed behavioral bookkeeping about one’s fellows. Finally, we will see in subsequent chapters that the most plausible model of group selection involves warfare and other forms of competition among groups. This mechanism would most likely favor feelings of solidarity and generosity towards the members of one’s group extending beyond relatives of recent common descent, accompanied by the lack of such feelings towards members of other groups. A common feature of all of these preferences is that helping is conditional: on close genetic relatedness, on the likelihood of future interactions, on the possibility of reputation building or signaling, or on group membership.

The fact that helping behaviors are indeed motivated by this wide range of proximate motives, from maternal love, to enlightened self-interest, to solidarity with one’s co-ethnics or co-nationals, is consistent with our view that in all likelihood each of the mechanisms we have identified here has played a significant role in human evolution.

Not surprisingly, given this account, recent (and therefore provisional) experiments in behavioral neuroscience have found that in some settings helping behaviors are correlated with activation in brain regions associated with cognitive operations such as the prefrontal cortex, as the reciprocal altruism, indirect reciprocity, and signaling models would predict, while in other settings brain activities associated with social reward processing are involved, consistent with models of kin altruism and multi-level selection.

All of the models in Table 6.1, and no doubt other mechanisms as well, were involved in the evolution of human cooperation, the importance of each depending on the forms of cooperation under consideration and the ecological and social conditions under which ancestral humans interacted. We will see in the next chapter that what can be known or conjectured about these ancestral human conditions from genetic, archaeological and
other data suggests that the two workhorse models of evolutionary biology, helping close family members and reciprocal altruism, do not singly or jointly provide an adequate account of the emergence of this cooperative species. In subsequent chapters we will show that group selection models based on gene culture coevolution contribute substantially to a convincing explanation.
Ancestral Human Society

[Among] prehistoric men... life was a continual free fight, and beyond the limited and temporary relations of the family, the Hobbesian war of each against all was the normal state of existence.

Thomas H. Huxley 1888.
The Struggle for Existence: A Programme, p. 163.

The philosophers... have all felt it necessary to go back to the state of nature, but none of them has succeeded in getting there.


Could strong reciprocity, fair-mindedness, and other altruistic behaviors documented in Chapter 5 be the legacy of an evolutionary past in which individuals behaving in these ways and their immediate families had higher fitness than other members of their groups. Trivers (2007) reasons that “unfair arrangements... may exact a very strong cost in inclusive fitness. In that sense, an attachment to fairness or justice is self-interested...(p. 77)” If Trivers is correct, fair-mindedness could have become common among humans if it benefited the individual or close relatives in repeated interactions, allowing fair-minded individuals to gain reputations that advanced their genetic interests. This, indeed, is the explanation of the evolution of fairness norms given by Ken Binmore (2005). The same reasoning may be applied to generosity and punishing those who transgress social norms. According to this view, Pleistocene conditions would have favored altruism among close family members and what we termed in Chapter 3 the “enlightened self-interest models” of cooperation.

Richard Dawkins (1976) advances exactly this view, having outlined kin altruism, reciprocal altruism, indirect reciprocity, and costly signaling as the four models likely to explain the origins of modern human altruism:

Throughout most of our prehistory, humans lived under conditions that would have strongly favored the evolution of
all four...most of your fellow band members would have been kin, more closely related to you than members of other bands...plenty of opportunities for kin altruism to evolve. And...you would tend to meet the same individuals again and again throughout your life-ideal conditions for the evolution of reciprocal altruism. Those were also the ideal conditions for building reputations for altruism and the very same ideal conditions for advertising conspicuous generosity.” (p. 220)

Thus, it could be that humans became cooperative because in our ancestral environments we interacted frequently with the same group of close kin, among whom tit-for-tat and other self-regarding strategies consistent with Trivers’ reciprocal altruism were sufficient to support cooperative outcomes. Cooperation thus emerged as a form of mutualism.

Others, including Leda Cosmides and John Tooby (1992) share this view of the evolutionary origins of social preferences, but in contrast to Trivers hold that in modern settings they are fitness-reducing, and hence maladaptive. Though the cooperation that was extended to family and reciprocating fellow group members enhanced the fitness of cooperators among our ancestors, according to Cosmides and Tooby’s interpretation, in the modern world of more ephemeral social contacts its expression is a maladaptive legacy of the distant evolutionary origins of human motivation. Because our ancestors rarely encountered strangers we just do not sufficiently distinguish, either in lab experiments or in real life, between one-shots and long lasting interactions, treating strangers much as if they were intimates. Dawkins explains it this way: “the lust to be generous and compassionate...is the misfired consequence of ancestral village life.” (p. 222)

There is little doubt that reputation-building in repeated interactions and a tendency of close genetic relatives to interact frequently contributed to the evolution of cooperation. But, we think it unlikely that these motivations are sufficient explanation. First, modern humans are perfectly capable of distinguishing between situations in which helping kin and maintaining a reputation are operative and situations when they are not, and it their prosociality in the is the latter situations, that accounts for the fact that humans cooperate non-repeated interactions and with unrelated others. Second, the view that early humans lived in worlds with little contact outside one’s family, Dawkins’ ideal conditions for self-interested cooperation to flourish, is difficult to square with what is known about the late Pleistocene and early Holocene. Like Jean-Jacques Rousseau’s philosophers, Richard Dawkins
and other modern-day biologists have jumped on a faulty time machine, and seem to have journeyed to an imaginary ancestral world.

The evidence for a quite different picture of our ancestral condition is necessarily indirect, but convincing. Pre-historic foragers left few archeological traces and the historical record contains few pre-contact histories extending over more than a half a century. The best we can do is to make inferences from the available data including hunter-gatherer demographics, late Pleistocene climate records, archeological evidence on causes of deaths during the Pleistocene, and ethnographic and historical reports on recent foragers.

We will see that neither the likely size of groups, nor the degree of genetic relatedness within groups, nor the typical demography of foraging bands is favorable to the view that late Pleistocene human cooperation can be adequately explained by kin altruism or reciprocal altruism. What is known or can reasonably be inferred about the Late Pleistocene and early Holocene suggests that ancestral humans did not live in small closed groups in which family and self-interest with a long time horizon alone were the cement of society. Rather our ancestors were cosmopolitan, civic minded, and warlike. They almost certainly benefitted from far flung co-insurance, trading, mating and other social networks, as well as from coalitions and, if successful, warfare with other groups.

Altruistic behaviors and the sentiments motivating them proliferated because ancestral humans both created social environments that partially shielded altruists from exploitation by the self-regarding and actively cultivated other-regarding preferences, modeled and simulated in Chapter 8 and Chapter 11 respectively, and because groups with many cooperative members were more likely to survive environmental crises and to prevail in lethal competition with other groups (Chapter 9).

In the next section we review the available archaeological and ethnographic evidence suggesting that while isolated groups surely existed, most humans had frequent contact with a substantial number of individuals beyond the immediate family. This conclusion is consistent with data on foragers reviewed in §7.2 on the extent of genetic differentiation among ethnographic foragers, by which we mean foragers in the historical and ethnographic record. Had ancestral groups been small and isolated, the extent of between-group genetic differences would have been considerably greater than is observed. We then provide evidence that ancestral humans engaged in frequent and exceptionally lethal intergroup conflicts. In the penultimate
section we turn to data suggesting that social order in pre-state small-scale societies was sustained in important measure by a process of coordinated peer pressures and punishment, a process that we will model and simulate in Chapter 10.

7.1 Cosmopolitan Ancestors

Were ancestral groups small enough so that cooperation could be supported by repeated interactions?

The minimal feasible foraging band, Christopher Boehm (2007) reasoned, would include five hunters. The number of adult decision-makers then would be triple or more this number, counting women and the elderly. The average band (census) size among the 175 ‘warm climate, non-equestrian’ hunter-gather groups identified by Frank Marlowe Marlowe (2005) as the groups in the ethnographic record most likely to be similar to ancestral humans is 37. Even if we exclude the old and the young, this would be about twelve adult decision makers. In Figure 3.3 and Figure 3.4 we showed that even for groups a third this size reciprocal altruism will evolve only for implausibly low rates of behavioral or perceptual error and extraordinarily high benefit-cost ratios of the altruistic behavior.

But the relevant number for human cooperation is not the average size of the elementary group of foragers. First, seasonal and other aggregations of many bands occurred for purpose of exchange, marital matching, sociality, and defense. Second, most ancestral humans necessarily lived in larger than average groups. For evolutionary explanation, either cultural or genetic, it is the size of the group that a typical individual lived in that is relevant, not the much smaller average group size. This so called “typical group size” may differ substantially from the average group size. Suppose there are two groups, one with 10 members and the other with 100 members. The average group size is 55. But the typical group size, weighting both groups’ size by their share of the total population, or $(100 \times 100 + 10 \times 10)/110$, is 92. Another way of seeing this is to note that 100 individuals have 99 neighbors and 10 individuals have 9 neighbors, so the average number of neighbors per individual is $(100 \times 99 + 10 \times 9)/110 \approx 91$; average group size is therefore $\approx 92$. Using this concept, we calculated the typical group size for the same “warm climate non-equestrian groups in Marlowe’s sample.” It was an astounding 77, more than twice the average group size. A group of this size would have 40 or more adult decision makers. We think even that
this is probably an underestimate, given that during the Late Pleistocene a far greater fraction of hunter-gatherers than today lived in large, partially sedentary villages in the relatively densely populated resource-rich coastal and riverine environments from which they were subsequently expelled by Holocene farmers.

More than limited group size is required to for repeated interactions to sustain cooperation: frequent and on-going interactions are also necessary (equation 3.2). But evidence of extensive contacts outside the group and cataclysmic demographic events make it unlikely that these requirements we met during the Late Pleistocene and early Holocene.

As we have seen in Chapter 1, long distance trade among foraging groups in Africa probably predates the diaspora of modern humans. Exchange was also common among Australian aboriginal groups:

Shell, stone, ochre, pituri [a narcotic] and wooden tools and weapons all circulated over hundreds of kilometers…shell pendants crossed the continent…A wooden hooked boomerang was seen…at least 1200 km. west of its area of manufacture. (Mulvaney 1976:80)

Aboriginal cultural practices also were exchanged. One, the Molongo ceremony, apparently traveled 1600 kilometers between 1893 and 1918 (Lourandos 1997:43)

Exchange often involved seasonal meetings of 1000 or more at which ‘communal food producing activities included…hunting drives [that] sometimes involved a human circle of 20 to 30 kilometers in diameter….These meetings were most often held close to territorial boundaries so as to avoid trespass and therefore further conflict (Lourandos 1997:65).

Nor were the aboriginal Australians atypical in this respect. Many groups of modern foragers meet in seasonal groupings of quite large aggregations of bands, constituting sometimes many thousands of individuals (Binford 2001, Kelly 1995). Exogamy also promoted contacts with strangers. The average distance between the birthplaces of husbands and wives among hunter-gatherers in tropical areas reported in MacDonald and Hewlett (1999) is 40 km, and much greater for some hunter-gatherers, such as the !Kung (70 km) and Arctic peoples (Fix 1999). Among the (‘pure’)}
!Kung groups studied by Harpending and Jenkins (1974) only 58 percent of the parents of the fertile adult members of the population were born in the same group as their children. These high levels of mobility, also noted in other Southern African groups such as the Hadza (Woodburn 1982), arise in part from the far-flung coinsurance networks, that among the !Kung may include partnerships well over 200 kilometers distant. For example, describing the likely historical patterns of the Dobe !Kung, Howell (2000):11 writes:

> The pattern of frequent long and short distance migrations, utilizing widespread kinship ties, must be an old pattern, with living groups forming and dissolving, coalescing and splitting to adjust groups size to the resources available. While no or few individuals may ever have covered the whole !Kung range during a lifetime, nearly every individual can be expected to travel long distances, living in a number of places and encountering a wide range of other !Kung, even if always in relatively small groups.

In these environments those who failed to distinguish between long-term and short-term or one-shot interactions would be at a significant fitness disadvantage as a result. It seems safe to conclude that all but the most isolated forager groups typically engaged in relationships, both beneficial and hostile, with other groups.

Turning to evidence on cataclysmic demographic events, frequent catastrophic mortality due to conflicts, environmental challenges and other causes is the most plausible way to reconcile two pieces of solid evidence about hunter-gatherer demography (Keckler 1997). First, human population grew extraordinarily slowly or not at all for the 100,000 years prior to 20,000 years before the present with estimated growth rates ranging from 0.002 percent per annum in the earlier period to 0.1 percent in the later (Bocquet-Appel et al. 2005). Models and data on hunter-gatherer demographies show that they are capable of growth rates in excess of 2 percent per annum (Birdsell 1957, Hassan 1980, Johansson and Horowitz 1986). The two facts are easily reconciled if population crashes were frequent.

Volatile climate must have contributed. Evidence of late Pleistocene temperature variability based on Greenland ice cores is presented in Figure 7.1. Deep sea cores in the Western Mediterranean and other data suggest that the climate variability shown in the figure was a general northern hemi-
Figure 7.1. Pleistocene temperature variations. Shown are measures of $\delta^{18}O$ taken from Greenland ice cores (from http://www.glaciology.gfy.ku.dk/ngrip/index_eng.htm and described in North Greenland Ice Core Project Members (2004)).

Sphere phenomenon (Martrat et al. 2004). Surface temperature scales approximately linearly with the $\delta^{18}O$ signal shown in the figure. Differences in temperature (Centigrade) are about 1.2 times the difference in the signal shown the figure (Johnsen et al. 1992). Even these data (smoothed to 50-year averages to reduce measurement error) indicate that changes in mean temperature as great as 8 degrees (C) occurred over time spans as short as two centuries. By way of comparison, the Little Ice Age that devastated parts of early modern Europe experienced a fall in average temperatures of one or two degrees, and the dramatic warming of the last century raised average temperatures by one degree, comparing the unprecedentedly hot 1990’s with a century earlier (Mann et al. 1998, McManus et al. 1999).
The high levels of mortality and frequent population crashes and dispersals that probably characterized the late Pleistocene are unfavorable to the evolution of reciprocal altruism, even in dyads. Paleo-demographic data suggest that the life expectancy at age fifteen during the late Paleolithic may have been about 17 years and a bit less for seven contemporary hunter-gatherer populations (Weiss 1973). Even the lower mortality hazard rates of contemporary foragers place an upper limit on the discount factor of 0.98, even in the absence of myopia, weakness of will, and the like (Gurven and Kaplan 2007) (the discount factor under these assumptions is 1 minus the probability of mortality). This purely demographic source of discounting raises doubts about results from repeated game theory that require the discount factor to be arbitrarily close to unity. If older members of the group are critical to the success of cooperation then the maximal discount factor will be considerably lower.

Moreover, due to frequent group conflicts and extraordinary climactic instability, group longevity was also probably quite limited, so that even if members of a group survived, they were unlikely to remain together over very long periods. Of fourteen groups studied in Papua New Guinea over a fifty year period by Soltis et al. (1995), five ceased to exist. The expected duration of both the Yanomamo villages reported in Smouse et al. (1981) and of the Gainj groups reported in Long et al. (1987) is about three generations. It is unlikely that groups survived longer than this under the turbulent climactic conditions of the late Pleistocene.

The only small scale society on which we have adequate long term demographic records, the people of the isolated South Atlantic island Tristan da Cunha, provide a dramatic example (Roberts 1971). Three population crashes occurred in the course of a century and a half. The first was a conflict-induced fission that resulted in the departure of more than half the population while the second, a sea-faring accident, killed all but four of the island’s male population. The final crash occurred with the dispersal of the entire population in 1961 due to an eruption of the island’s volcano.

Summing up a pattern of demographic instability likely to have been faced by the foraging bands of the late Pleistocene, Gajdusek (1964) wrote:

Migrations, murders and suicides, warfare, and massacres, splitting and fragmentation of communities and bands, or amalgamations and sudden mergers of groups, sudden social changes in mating practices and prohibitions and such natural accidents and catastrophes as earthquakes, floods, typhoons,
volcanic eruptions, droughts, famines, and plagues all have a major influence in determining the genetic composition of small groups. (p. 121)

These are hardly conditions under which reciprocal altruism could be expected to flourish (Gintis 2000). Genetic evidence is consistent with this view.

### 7.2 Genetic Evidence

The extent of genetic differentiation among groups of human foragers is the result of mating practices, selection pressures, reproductive inequalities, migratory patterns, group size, the process of group fission and fusion, and other aspects of population dynamics over the very long run. While contemporary foraging populations are not pristine replicas of ancestral human groups, genetic evidence on foragers nonetheless provides a rare lens with which to study the pre-history of human social and demographic structure.

Genetic material collected from foraging populations over the past half-century allows estimates of the degree of genetic distance between groups on scales ranging from elementary foraging bands to meta-populations composed of many ethno-linguistic units. These estimates, taken from Bowles (2006), augmented by more recent data appear in Table 7.1.

Many foraging populations are highly differentiated. Some geographically adjacent groups are as different genetically from one another as are the major ancestral groups of the world. Foraging populations do not appear to be less differentiated genetically than horticultural and more technologically advanced small-scale populations. But, our joint work with Stefany Moreno Gamez and Jon Wilkins of the Santa Fe Institute indicates that if ancestral groups were indeed small and closed, the degree of differentiation among groups predicted by the standard model of equilibrium differentiation would be substantially greater than what we observe (Gamez et al. 2010). The genetic data thus are more consistent with ancestral groups being of considerable size and with ever-changing composition.

Equilibrium genetic differentiation balances the effects of small group size and the resulting genetic drift that tends to enhance differentiation on the one hand and between-group migration, tending to reduce it on the other. The equilibrium level of differentiation, denoted $F_{ST}$, is that at which these two effects are balanced, leading to long run constancy of the $F_{ST}$ in the
absence of exogenous changes in the underlying data. For autosomal markers, assuming that the alleles in question are not subject to selection or mutation and letting $m_e$ and $N_e$ be measures respectively of the rate of migration among groups and group size we have the following approximation for small $m_e$ (Wright 1935):

$$F_{ST}^* = \frac{1}{1 + 4m_eN_e}.$$  \hspace{1cm} (7.1)

Variance effective group size ($N_e$) is the number of individuals in a single generation in an idealized group (whose features are described below) that would result in the same amount of genetic change due to genetic drift as would the empirical group under study, given its demographic structure when account is taken of its deviations from the idealized structure. The idealized population is constant, and the expected number of progeny contributed to the next generation are equal among both males and females, who are also equal in number. The effective migration rate ($m_e$) is based on the idealized random island model, where migrants are assigned randomly to groups independent of spatial or other proximity, and is the rate that would have the same effect on the degree of equilibrium differentiation as the observed migratory structure. No human societies exhibit these idealized features and the level of genetic differentiation predicted for observed demographic structures is greater than under idealized conditions.

Before turning to a more realistic model, first note that under these idealized assumptions the observed measures of differentiation are consistent with equilibrium predictions for groups of modest size and migration rates. Let $m$ be the observed migration rate, namely the fraction of the population that relocates to another group per generation and let $N$ be the number of deme members of a breeding generation, which is about a third of the census size (Cavalli-Sforza and Bodmer 1970, p. 416) because approximately one generation in three is of reproductive age, but less in some cases (Rychkov and Sheremet’eva 1980, p. 49). If the census group size is 37, we have a breeding generation size, equal for males and females, of $N \approx 12.3$. According to equation 7.1 the median observed $F$-statistic from Table 7.1, 0.076, is an equilibrium value if demes averaged three migrants per generation, namely $m_eN_e = 3$. This is because $0.076 = 1/(1 + 4 \times 3)$. We write this number as $M^*$ and use it as a composite measure of group size and migration rates consistent with the observed $F$-values. Thus if ancestral groups were of the size that Marlowe suggests, a migration rate between
groups of 0.24 would yield 0.076 as the predicted $F$-value, very close to the mean of 0.08 in Table 7.1.

But, it is unlikely that the idealized assumptions underlying equation 7.1 obtained during the Pleistocene. Reproductive skew, volatile population dynamics of the type documented above, including rare group extinctions, lineage-based group fission, and non-random migration, all must have been common during the Late Pleistocene. All are known to contribute to elevated equilibrium genetic differentiation among groups. Using data on recent foraging populations to take account of these effects, predictions from the equilibrium model are broadly consistent with the data, but, as we will see, only if group size and migration are substantial. Thus, if observed genetic differentiation reflects a population equilibrium, the data are inconsistent with the common view that our ancestors lived in closed worlds in which social interactions were confined to a small number of relatives or lifelong close associates.

In collaboration with Stefany Moreno Gámez and Jon Wilkins, we computed the implications for likely group size and migration rates of substituting a more realistic set of assumptions for the idealized model underlying equation 7.1. We simulated the evolution of between-group genetic differentiation in a large population with four empirically based characteristics. We designed a mating system so that within-group reproductive skew approximated that observed in modern foraging populations. We simulated population crashes and extinctions on a scale likely to have occurred during the Pleistocene. When new groups were formed, either to populate a site vacated by an extinct group or because a group fissioned, we populated the new site, or the breakaway group, with members of a single parent group rather than by a random draw from the entire population. Finally, we replaced the random migration model with a tendency of individuals to relocate in groups where their relatives live.

We found that taking account of these deviations from the idealized model, reproductive skew, population crashes, non-random re-colonization and migration, we found that the number of migrants per year sufficient to yield the mean of the observed $F_{ST}$’s rose from 3 to 8.

Eight migrants per generation could occur, leaving the migration rate unchanged at 24 percent, if the size of a single reproducing population were 33 rather than 12, implying a census size of about one hundred. If group size remained unchanged, at 12 per generation, the observed $F_{ST}$ would be an equilibrium only if two-thirds of the group migrated each generation.
Thus the genetic evidence suggests that ancestral groups were either quite large or experienced high levels of migration, or both.

More detailed studies suggest a similar conclusion. Based on the pattern of genetic differentiation among aboriginal Australians groups, for example, Keats concludes that “local groups traveled large distances and often came into contact with one another for the purposes of trade, which sometimes included exchange of people from each group” (Keats 1977, p. 327). Inferences from these genetic data are thus consistent with evidence of both long-distance exchange (McBrearty and Brooks 2000) and hostile conflict (Bowles 2009a) among ancestral groups, and among recent foragers thought to be plausible models of ancestral humans, significant levels of long-distance exchange and migration, occasional large scale seasonal agglomerations, and far flung co-insurance partnerships (Lourandos 1997, Wiessner 2002, Marlowe 2005).

These data thus are consistent with the above evidence that ancestral conditions were not favorable for the evolution of group level cooperation by means of reciprocal altruism. The genetic data also support a similarly negative judgment for the adequacy of the kin selection model. Members of groups were not very closely related.
Table 7.1. **Genetic Differentiation among Hunter-gatherer Populations.** $F_{DG}$ measures genetic differentiation among demes (D) in the same ethno-linguistic group (G), while $F_{GT}$, and $F_{DT}$ respectively measure differentiation among groups and demes in a meta-population (T). The mean value of the augmented data set is 0.080 (0.041) and the median is 0.075. If within-ethno-linguistic unit estimates ($F_{DG}$) are excluded, the remaining 15 estimates have a mean of 0.087 (0.038) and a median of 0.076. The Pygmy and Arnhem-North Australian data are based on microsatellite loci corrected to be comparable with the remainder of the table. The first 15 rows are from Bowles (2006). These data and the corrections are described in a memo at http://www.santafe.edu/~bowles. The next two (Pygmy) are from Verdu et al. (2009), the next (Arnhem and Northern) are from Walsh et al. (2007) and the final three rows are from Cavalli-sforza, Menozzi, and Piazza (1994).
Though helping siblings and other close family members would have been supported for plausible benefit cost ratios, relatedness within typical foraging bands would not support cooperation among band members generally unless the benefit cost ratio is extraordinarily high. Consider a concrete example of an exceptionally highly related band (of just seven adult members) with two parents and their two sons and wives and one unrelated ‘outsider’. If the spouses are unrelated, the average pair-wise degree of genetic relatedness in this group is 0.12. Hamilton’s rule (3.1) shows that kin altruism would be supported among group members only if the benefit cost ratio were to exceed eight (i.e., the inverse of 12). The estimates of the fraction of the total genetic variance that is between demes reported in Table 7.1 suggests an average degree of relatedness in demes that would preclude the evolution of altruistic behaviors unless the benefits were an order of magnitude greater than the costs. Of course the same data present a challenge to any model of the evolution of altruistic behaviors based on positive genetic assortment, including multi-level selection models. In the next chapter, we will see that between-group competition may have provided the necessary extraordinary benefits of group level cooperation. The main reason why this is so is that conflict among hunter-gatherer groups was common and exceptionally lethal.

### 7.3 Prehistoric Warfare

Not withstanding a number of insightful recent studies (Burch 2005, Gat 2006, LeBlanc 2003), lethal inter-group conflict among hunter-gatherers during the Late Pleistocene and early Holocene remains a controversial and, like the question of altruism itself, ideologically freighted subject, with little agreement on either its extent or consequences (Ferguson 1997, Keeley 1996).

Among the empirical challenges are the lack of written accounts, the difficulty in making inferences from ethnographic hunter-gatherers in the ethnographic record about conditions before the domestication of plants and animals and the emergence of states, and the fact that most foragers made little use of fortifications and killed each other with the same weapons that they used to hunt other animals, thus leaving few distinctive archeological traces other than skeletal remains.

The virtual absence of archaeological evidence of persistent economic and political differentiation between families prior to about twenty-four
Ancestral Human Society 137

thousand years ago indicates that the most informative data for understanding Late Pleistocene and early Holocene humans pertain to hunting and gathering populations without formal political structures (chiefs, big men, or states). We exclude populations making substantial use of domesticated plants and animals, namely, pastoral, horticultural, agricultural or equestrian hunting populations. As hunter-gatherer populations occupying resource-rich areas in the Late Pleistocene and early Holocene were probably sedentary (at least seasonally), we have included wars involving settled as well as purely mobile populations.

By ‘wars’ we mean events in which coalitions of members of a group seek to inflict bodily harm on one or more members of another group. The term is not ideal for the ambushes, revenge murders, and other kinds of hostilities likely to have occurred between ancestral groups of humans. Most hostile inter-group contact among hunter-gatherers was probably on-going or intermittent, with occasional casualties, more akin to boundary conflicts among chimpanzees (Manson and Wrangham 1991) than to modern warfare. However ‘pitched battles’ did occur among foragers, as in the conflict between two coalitions of aboriginal Australians involving around 700 combatants (Lourandos 1997, 59,61).

Using these definitions and selection criteria, we studied all available archaeological and ethnographic sources that present, or are cited as presenting, relevant data (Table 7.2). Of these thirty-four sources, fourteen were found to present data that was unrepresentative (for example when warfare was primarily with modern agricultural populations), unreliable, or inadequate. In three cases re-estimation of the critical information was possible. Skeletal evidence from sites with fewer than ten individuals was also excluded. Possible biases in this data set are discussed below. The eight ethnographic and fifteen archaeological included sources yield similar results, consistent with the view that prehistoric warfare was frequent and lethal, but somewhat less so than estimates based on data in the standard source for these estimates (Keeley 1996). The populations studied appear in Figure 7.2. Details and additional caveats these and the data to follow appear in (Bowles 2009b).

As with all archaeological data, it is difficult to establish if the sites that have been studied are representative of Late Pleistocene and early Holocene conditions. Most of the archaeological data are based on evidence of violent death such as spear points embedded in bones. As this evidence comes from burials, it is almost certainly not representative in one respect: sim-
Fig. 7.2. **Between-group Differences.** Sources of archaeological (black squares) and ethnographic (black closed dots) evidence on warfare and genetic (black open dots) data on between-group differences.

<table>
<thead>
<tr>
<th></th>
<th>Weighted Mean</th>
<th>Arithmetic Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaeological</td>
<td>0.12</td>
<td>0.14</td>
<td>0.12</td>
</tr>
<tr>
<td>Ethnographic</td>
<td>0.16</td>
<td>0.14</td>
<td>0.13–0.15</td>
</tr>
<tr>
<td>All</td>
<td>0.14</td>
<td>0.14</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Table 7.2. **Fraction of total mortality due to warfare (δ): summary statistics.**

Note: complete sources, methods and other details for this and Table 7.3 are in Bowles (2009a) and Bowles (2009b). Weights are the square root of the total number of deaths.

The disposal of the dead (rather than burial) appears to be typical of the archetypal so-called “immediate return” foraging group (Woodburn 1982). There may be more than accidental bias in the burials studied for signs of violence, given that evidence of violent deaths may be deemed more interesting or worthy of publication than the absence of such evidence. Evidence on given individuals are also incomplete, leading to the opposite bias. Most skeletal remains are never found, and those that are range from intact to fragmentary or poorly preserved, often comprising just a few of the 100 or so bones in an adult human, setting aside the small bones of the hands and feet. The remains of 2185 pre-historic people of present-day California are accessible to researchers in a museum collection that totals only 12,044 bones, excluding hands and feet; more than ninety percent of the individuals’ bones are absent (Tenny 1990).
Moreover, while some osteological evidence is indicative of ongoing inter-group violence (simultaneous burials, severed limbs and other evidence of trophy taking, for example), one cannot always distinguish between deaths due to inter-group violence and that occurring within groups. Other biases may lead to underestimates. Many deaths in warfare do not leave projectile points embedded in bone or the other traces of violent death: “an analysis that included only projectile points embedded in bone would miss over half of the projectiles…and 75 percent of what was in all probability the actual number of projectile wounds” (Lambert 1997). Studies of arrow wounds treated by U.S. Army surgeons during the Indian Wars found that less than a third of the arrows struck bone (Milner 2005) and that 61 percent of fatal arrow wounds were to the abdomen (Bill 1862). Finally, fatalities during combat may fall far short of the total effect of warfare when account is taken of the mortality and reduced reproductive success occasioned by the displacement of the surviving losers. Table 7.3 gives the resulting estimates.

Table 7.3. Archaeological evidence on the fraction ($\delta$) of adult mortality due to warfare. Note: before present indicates before 2008 (Bowles 2009a).
Table 7.4. Ethnographic evidence on the fraction (δ) of adult mortality due to warfare. Note: a indicates that the group were foragers; b maritime; c seasonal forager-horticulturalists; d sedentary hunter-gatherers; e recently settled (Bowles 2009a, Bowles 2009b).

Most ethnographic studies of pre-modern war have concerned populations whose unusually bellicose relations among groups may not reflect conditions of Late Pleistocene hunter-gatherers: horticultural peoples in the highlands of Papua New Guinea and parts of lowland South America, or equestrian hunters or sedentary horticulturalists in North America. Among non-equestrian foragers, detailed accounts provide examples of inter-group conflict of exceptional brutality among Aboriginal Australians, Eskimos and other groups (Burch 2005, Melbye and Fairgrieve 1994, Morgan 1979[1852]), but most do not allow quantitative estimates of the resulting mortality. In other groups war is entirely absent from the ethnographic record, but in some of these cases, like the !Kung and other Southern African groups this may be the result of recent state interventions (Campbell 1986, Schapera 1930). For eight populations ethnographic studies allow estimates of the deaths due to warfare as a fraction of total mortality (summarized in Table 7.4). As in the case of archaeological studies, selection bias may lead to an exaggeration of the extent of warfare mortality. Moreover, some populations are not entirely representative of foragers during the Late Pleistocene due to the impact of non-hunter-gatherer influences.

The mortality data summarized in Table 7.2 are consistent with what is known about the Late Pleistocene from more indirect data, for example, as we have seen, reconciling the capacity of forager populations to expand rapidly coupled with the virtual absence of population growth until the very
end of the Pleistocene. Further, the extraordinary volatility of climate during the Late Pleistocene must have resulted in natural disasters and periodic resource scarcities, known strong predictors of group conflict among hunter-gatherers in the historical record (Ember and Ember 1992), and undoubtedly forced long-distance migrations and occasioned frequent encounters between groups having no established political relations. In light of the climate record Boehm (2000):19 writes that:

...towards the end of the Pleistocene as anatomically modern humans began to emerge, group extinction rates could have risen dramatically as needy bands of well-armed hunters, strangers lacking established patterns of political interaction frequently collided, either locally or in the course of long-distance migration.

A statistical analysis of recent ethnographic evidence is consistent with Boehm’s conjecture. Ember and Ember (1992) conclude that a “history of unpredictable natural disasters strongly predicts more war...people, particularly in non-state societies, may try to protect themselves against future disasters by going to war to take resources from enemies.”

The impact of climate variation is also suggested in the archaeological record. Commenting on a burial from 12 to 14 thousand years ago in which almost half of the skeletons indicated a violent death, Wendorf (1968):993 explained:

Population pressures may have become too great with the deterioration of Late Pleistocene climate and the effects which this had on the herds of large savanna type animals which were the primary source of food at this time. ...a few localities which were particularly favorable for fishing would have been repeatedly fought over as sources of food became increasingly scarce.

Archaeological evidence on Southern Californian maritime hunter-gatherers over a 7000 year period indicates that violent deaths occurred disproportionately during periods of climatic adversity and resource stress (Lambert 1997).
Social interactions within ancestral groups, we believe, exhibited an interplay of conflict and cooperation not dissimilar from between-group interactions. A difference however is that aggression within groups often served to sustain cooperation. The alacrity with which experimental subjects punish defectors and the extent to which people in modern natural settings will go to shun, ridicule, humiliate and harm those who violate social norms have their counterparts in small-scale societies. We think that, like intergroup warfare, these forms of peer punishment are both an exemplar of altruistic cooperation and one of the practices that among our ancestors curbed self-aggrandizement and provided a favorable environment for the evolution of cooperative behaviors. In subsequent chapters we will explain how warlike predispositions and the willingness to bear costs to punish defectors in one’s own group could have proliferated in the social and ecological environment of the late Pleistocene and early Holocene. But to do this we must, as in the case of warfare, know the nature of the behaviors in question.

Balikci (1970) reports that among the Netsilik, an isolated hunting people on the Arctic coast,

…there is a general rule…according to which all able bodied men should contribute to hunting, and the returns of the hunt should be shared according to established custom. Any activity in exception to this rule was bound to provoke criticism, various forms of conflict, and frequently social ostracism. (176)…lazy hunters were barely tolerated by the community. They were the objects of back biting and ostracism…until the opportunity came for an open quarrel. Stingy men who shared in a niggardly manner were treated similarly. (177)

In like fashion, Richard Lee (1979):458 describes the moral world of the !Kung in Southern Africa:

The most serious accusations one !Kung can level against another are the charge of stinginess and the charge of arrogance. To be stingy, or far-hearted, is to hoard one’s goods jealously and secretly, guarding them “like a hyena.” The corrective for this is to make the hoarder give “till it hurts”; that is to make him give generously and without stint until everyone can see that he is truly cleaned out. In order to ensure compliance with
this cardinal rule the !Kung browbeat each other constantly to be more generous and not to hoard.

Most of the lethal violence documented by Lee concerns such problems as adultery, not stinginess. By contrast to the reports of Lee and Balikci, Endicott (1988:118 reports horror expressed by a Batek informant at the thought of exiling a member whose laziness had caused some resentment.

But there is no reason to doubt the thrust of their observations, that those who violate a social norm by hoarding or slacking would be in for a rough time. Christopher Boehm’s (2011) survey of social control methods in what he calls “Pleistocene-style” hunter-gatherer populations shows just how rough it can get. Among the three hundred or so forager societies described in ethnographies, he selected those likely to match the social and environmental conditions of Late Pleistocene humans. This involved eliminating mounted hunters, bands dependent on religious missions or that had extensive trade with farmers, or had engaged for centuries in the fur trade, or that made significant use of stored food. He then scoured all available accounts for evidence of transgressions of social norms, and how they were addressed in these small groups lacking both police and courts of law. Stealing and murder were condemned in all of the societies, and “failure to share” was considered a violation of a norm in 43 of the 53 societies studied, the same number that condemned “beating someone”. “Bullying” was frowned on in 34 of the societies. The common responses in most societies in response to these transgressions was distancing, shunning, ridicule, shaming and gossip. In twenty of the societies expulsion from the group occurred, and in striking thirty-four of the populations, assassination of the culprit by the entire group was practiced. Boehm’s conclusion: “yesterday’s hunter-gatherers were well equipped to identify free-riders, suppress their behavior, and—if they couldn’t intimidate them enough to keep them under fairly good control—get rid of them.” Punishment often goes considerably beyond browbeating.

The only quantitative investigation of social order in a pre-state society of which we are aware is remarkable study by Wiessner (2005). During 1974 and 1996-7, she recorded 308 three- to four-hour time blocks of conversations among the Ju/'hoansi (!Kung) of the Dobe-/Kae/kae area in Northwest Botswana. In the 1970s, the Ju/'hoansi obtained virtually all of their subsistence through hunting and gathering. By the 1990s their society had been transformed, with government rations, wages, the sale of crafts, and old age pensions occupying a substantial role in their subsistence. Of the conversa-
tions studied, 56% included criticism of one or more members of the group including harsh criticism in two fifths of the cases, and ridicule and mild forms of expression of displeasure in a fifth of the cases. The remaining two fifths of the cases were explicit but not harsh criticism. Only 7 percent of the conversations included praise. Frequently, targeted behaviors included stinginess, greed or laziness, including reclusive behavior associated with failure to share, big-shot behavior, failure to honor kin obligations, and drunkenness. The intent of criticism was behavior modification. In sixty-nine percent of the cases the target was either present or within earshot. In 21 percent of the cases the target was absent but his close relatives were present. The remaining cases involved coalition building for punishment at a later date.

Wiessner recounts one of these episodes: “In 1974 a highly respected leader and one of the few cattle owners was accused of big-shot behavior and failure to share meat. After suffering extensive criticism in silence, he noted that one of his cows was behaving dangerously, slaughtered it, and distributed the meat widely.” Another “big-shot” was initially subjected to ridicule, but when this proved ineffective, harsh criticism followed along with a refusal to share meat with the offender, who eventually left the group for a period. A disruptive woman was ostracized from the group and subsequently died, after which her family was welcomed back into good standing. In one case criticism aimed at expulsion of a segment of the group led its relocation at a new camp 9km distant, after which friendly relations were resumed.

The content and targeting of these conversations did not differ between the 1970s and 1990s. Men and women initiated criticism about equally, but men were overwhelmingly the targets of criticism because they more frequently engaged in big-shot or disruptive behaviors. High status (“strong”) group members engaged in more criticism than others, and were as likely to be targeted as “weak” members. The most respected figure in the camp (the n'ore kxao) tended to refrain from criticism, thereby facilitating reconciliation so as to avoid the loss of a group member.

Four aspects of social order among the Ju/'hoansi are important in what follows. First, though Wiessner stresses the many ways that the Ju/'hoansi reduce the costs of maintaining order, those engaging in norm enforcement bore costs in the form of disrupted patterns of sharing. In 3% of the cases violent brawls resulted. Second, most of the criticisms were carried out by coalitions (three or more persons) rather than individuals, and this was
true in virtually all of the cases where the target had threatened community
stability or harmony. As a result, the costs of engaging in punishment of
norm transgressors was relatively low as long as long as the number of
altruistic punishers was sufficiently great. In Chapter 10 we will explain
how this kind of coalitional punishment of transgressors may have evolved.
Third, the punishment of those who refrained from punishing was entirely
absent: “I have not observed any second order punishment in all of my
years working with the Ju/'hoansi.” Fourth, Wiessner finds no evidence
that punishing norm transgressors is a signal of otherwise unobservable
quality as a mate, coalition partner, or competitor. Single reproductive-
age individuals are the least likely to punish, contrary to the predictions of
the signaling model. Moreover, while mediation ability is highly valued,
those who punished frequently or harshly were not sought out as coalition
partners but were considered to be tchi n!ai, “angry, sharp or biting things,”
and were told to desist.

The other study of the maintenance of order in a pre-state society that
has guided our attempts to model the evolution of cooperation concerns
ostracism among the Pathan Hill tribes an acephalous, egalitarian kinship-
based people in Afghanistan. Niloufer Mahdi (1986) writes:

The most striking use of ostracism among the Pathans is the
rejection by the tribe or clan of one of its members whose beh-
avior might lead to a feud...[O]stracism functions...to deter
behavior that violates customary legal norms, to punish spe-
cific acts that are culturally defined as improper, and to unify
the primary reference group on which individuals depend for
protection and economic support. (p. 295)

The Pathans, “do not recognize permanently established or permanently
functioning authority either in the central government or at any level of the
tribal structure.” (p. 296) Rather, social order is sustained by peer pun-
ishment. Ostracism is commonly the fate of those who have violated the
Pukhtunwali (Code of the Pathans).

The obligation to the Code is not secured in the first instance by
any coercive force. Ask any Pathan why he upholds Pukhtun-
wal and the answer will be because of izzat (honor). Obedi-
ence to Pukhtunwali is freely given. 297
Similarly, ostracizing those who have violated the Code “becomes the obligation of every Pathan, acting individually or as part of a relevant tribal segment.” (p. 297) Punishment of those who violate social norms is itself a norm.

Ostracism, writes Mahdi, may include “avoidance, exclusion from social participation, shunning or derision.” (295). If a man has to be avoided, he may not be considered for a matrimonial alliance or invited to participate in a Jirga (Council of Elders).” (p. 302) Notably, the aim of ostracism seems to be punishment, not behavior modification:

the unique expression of overt verbal disapproval that is practiced is paighor (insult or taunt). Its main purpose is to shame the person. It is most commonly applied in those cases where a man has been perceived to have failed in valor or in the discharge of an obligation of honor. Paighor is not given with the conscious aim of prompting its object to action, nor is it calculated to reform. However very often it has precisely that effect. (p. 303)

As in Boehm’s survey (1993), punishment may take the lethal forms, such a badal (revenge killing) or physical expulsion, the consequences

    when a Pathan is exiled from his group his situation becomes untenable. He is vulnerable not only to his dushman (the one who has targeted him), but can be victimized by anyone without fear of badal. (p. 301)

The fact that badal applies equally to all members of one’s kin group provides strong incentives for discipline to be applied within families, therefore reducing the frequency of more costly between-kin group feuding (as in the “community responsibility” model of Greif, 2002).

As among the !Kung, the egalitarianism of the Pathan peoples is essential to the process of peer punishment, as is the on-going nature of social interaction:

    …neither economic or social status deflects its course…if a man seeking badal is weak vis a vis his dushman he will pass on his obligation to his sons and they in turn to their sons. And if a man is well protected enough to escape badal himself, it is extremely doubtful that the protection can be extended to his
kin or to successive generations, who would constitute legitimate targets of badal. (p. 298)

Like the Ju/'hoansi, the Pathans attempted to minimize the costs of punishment. “To mitigate the disadvantages of badal, Pukhtunwali makes provision for mediation through a Jirga... when it appears that honor has been satisfied and it is an opportune time to halt the cycle of violence... Certain conditions such as monetary compensation may be imposed to achieve a balance in the redress of grievances.” (p. 299)

7.5 The Crucible of Cooperation

It has been conventional since Thomas Hobbes’ Leviathan to attribute the maintenance of social order to states. But, for at least 95 percent of the time that biologically modern humans have existed, our ancestors somehow fashioned a system of governance that without the assistance of governments avoided the chaos of the Hobbesian state of nature sufficiently to become by far the most enduring of social orders ever. The genetic, archaeological, ethnographic, and demographic data make it quite clear that they did not accomplish this by limiting human interactions to a few close genetic relatives. Nor can the oft-repeated nature of their interactions explain this remarkable accomplishment. Rather we will show in the next three chapters that a particular form of altruism, often hostile towards outsiders, punishing towards insiders who violate norms, coevolved with a set of institutions—sharing food and making war are examples—that at once protected a group’s altruistic members and made group-level cooperation the sine qua non of survival.
The Coevolution of Institutions and Behaviors

A population can be very successful in spite of a surprising diversion of time and energy into aggressive displays, squabbling and outright fights. The examples range from bumble bees to European nations.


p. 345

Few students of human social dynamics doubt that institutions, nations, firms, bands, and other groups are subject to selective pressures. The emergence and diffusion of the centralized, tax-collecting and arms-bearing national state as a form of territorial governance during the past half millennium is an example. The national state became the dominant form of governance because it won wars and induced preemptive emulation among those threatened with military subjugation (Tilly 1975, Bowles and Gintis 1984, Bowles 2004). Similar arguments may explain the evolutionary success of other social arrangements, markets, monogamy, private property, worshiping supernatural beings, social ranking, and sharing the necessities of life among non-kin, for example. These have been ubiquitous over long periods of human history and have emerged and persisted in highly varied environments. Talcott Parsons (1964) termed these arrangements evolutionary universals, the most likely explanation of which is that societies adopting them prevailed in competition with other groups. Frederich Hayek (1988) referred to the markets and private property nexus, his “extended order,” in a similar vein, attributing its success to cultural group selection.

Group competition and culturally transmitted group differences in institutional structure are central to our explanation of the evolution of cooperative behaviors among humans. We stress intergroup competition for empirical reasons: group conflict and the extinction or subjugation of loser populations have been among the most powerful forces contributing to the emer-
gence, proliferation and persistence of novel human behaviors and institutions (Parsons 1964, Tilly 1981, Bowles 2009b). The institutions we model are the commonly observed human practices of resource sharing among group members beyond the immediate family.

Group differences in institutional structure persist over long periods of time due to the nature of institutions as conventions. A convention is a common practice that is adhered to by virtually all group members because the relevant behaviors, for example sharing meat, or not engaging in extra-pair copulations, are mutual best responses conditional on the expectation of similar behaviors by most others (Lewis 1969, Young 1995). We do not here model the reasons why the behavior prescribed by the institution is a mutual best response, but plausible accounts are not difficult to provide (Kaplan and Gurven 2005). Individuals approve of sharing norms because they are beneficial to group members, and as we saw in the previous chapter, those violating these norms may bear fitness costs of ostracism and shunning, for example Boehm (1993). The conventional nature of institutions accounts for their long term persistence and also their occasional rapid demise under the influence of shocks.

The inheritance of group-level institutions results from a cultural transmission process based on learned behaviors. While copying behaviors that are common in a population is a strong influence on learning independently of the associated payoffs, we simplify by abstracting from this so-called conformist cultural transmission. Thus in our models when new members of the population mature or immigrate, they adhere to the existing institutions, not due to any conformist predisposition, but because this is a best response as long as most others do the same. The resulting behavioral uniformity in adherence to a group’s institutions permits us to treat the institution as a group-level characteristic. By contrast, the group-beneficial individual traits in our model are replicated by a standard fitness-based mechanism in which the above best response dynamic producing within-group uniformity are absent.

We study institutional evolution in ways analogous to the evolution of individual traits (§2.3). Just as the individuals in our model are the bearers of genes or socially learned individual behaviors, groups are the bearers of institutions, and a successful institution produces many replicas, while unsuccessful ones pass out of existence without leaving a trace. Replication of institutions may take place when a successful group grows and subdivides, forming two groups, or when a group with unsuccessful institutions
succumbs to a military, ecological or other challenge and its vacated site is occupied by colonists from a neighboring group.

The evolutionary mechanisms involved in this account are multi-level selection processes with the novel features that both genetically transmitted influences on individual behaviors and culturally transmitted group-level institutional characteristics are subject to selection. The model is thus an example of a gene-culture evolutionary process introduced in Chapter 2, with institutions playing the role of culturally transmitted niches, constructed environments affecting the processes of selection acting on genetically transmitted traits (Odling-Smee et al. 2003). We will show that these niches may allow a genetically transmitted altruistic predisposition to emerge when rare and then proliferate, and the resulting altruistic behaviors contribute to the competitive survival of those groups in which these niches have been constructed.

This chapter departs in three ways from the material presented thus far. The first, already mentioned, is our introduction of culturally transmitted group institutions and their analysis in a gene-culture coevolutionary model. Second, we use a group selection model based on selective extinction, and show that it provides a much more plausible account of the evolution of altruism than does the selective emigration model considered in §6.1. Finally, the plausibility of group selection models is an empirical question, so we put our model to an explicit empirical test. We ask: on the basis of archeological, genetic, and other data presented in the previous chapter, is it likely that genetic differences among groups were great enough and that the survival advantages of predominantly altruistic groups were a sufficiently powerful evolutionary force that under the conditions that ancestral humans experienced during the Late Pleistocene and early Holocene, altruism could have proliferated in the way we suggest? Readers interested in more extensive treatments of the empirical estimates and simulations reported here may wish to consult Bowles (2006,2007,2009a) and Bowles et al. (2003). Readers unfamiliar with multi-level selection models may wish to consult §6.1.

8.1 Within-group Variance Reduction

Evolutionary processes are strongly influenced by mechanisms that suppress competition and reduce differences in behavior or other phenotypic variance within entities, whether they be multicellular organisms or groups
of ancestral humans. In a paper studying slime mould (*Dictyostelium discoideum*) Steven Frank Frank (1995) writes: “Evolutionary theory has not explained how competition among lower level units is suppressed in the formation of higher-level evolutionary units,” (p. 520) adding that “mutual policing and enforcement of reproductive fairness are also required for the evolution of increasing social complexity.” John Maynard Smith and Eors Szathmary note that many of what they term the “major transitions in evolution” share a common feature: “entities that were capable of independent replication before the transition can replicate as part of a larger whole after it.” (Maynard Smith and Szathmáry 1997), p. 6. As a result, the constituent entities making up the higher-level units come to share a common fate, with selection pressures working on the higher- rather than the lower-level units.

Christopher Boehm (1982):421 applied this idea to human evolution. Among ancestral foragers, he wrote, “group sanction emerged as the most powerful instrument for regulation of individually assertive behaviors, particularly those which obviously disrupted cooperation or disturbed social equilibrium needed for group stability.” As a result (Boehm 1999):211: “a ‘political revolution’ experienced by Paleolithic humans created the social conditions under which group selection could robustly support genes that were altruistic.” Relatedly, Irena Eibl-eibesfeldt (1982):177 pointed to the importance of “indoctrinability to identify with values, to obey authority, and. . . ethical sharing” and thought that “through these bonding patterns, groups become so tightly knit that they could act as units of selection.”

As a result, characteristics of groups such as the institutions that regulate the selective pressures among their members become themselves subject to selective pressures. An example of such a group characteristic, one stressed by Boehm, and we will explore here, are leveling institutions such as food sharing beyond the family, namely those which reduce within-group differences in material well-being. By reducing within-group differences, such structures would have attenuated within-group selective pressures operating against individually costly but group-beneficial behaviors, so that groups adopting these leveling institutions retained a greater number of altruists, thus conferring advantages for these groups. In this case the ubiquity of group structural characteristics such as leveling institutions is explained by their contribution to the proliferation of group-beneficial individual traits and the contribution of these traits to group survival when faced with environmental crises or competition from other groups.
The individual behaviors that we consider include warning others of danger, acquiring and sharing valuable information, participation in the defense of the group or in predation of others, or the punishment of those who fail to conform to these group-beneficial behaviors (Clutton-Brock and Parker 1995, Boyd et al. 2003). Thus, the formally altruistic (individually costly but in-group beneficial) traits that may proliferate under the influence of multi-level selection include behaviors that may be harmful to members of other groups. The processes modeled here might be best described, paraphrasing Laland et al. (2000):224, as demonstrating the evolutionary success of selfish groups rather than generous individuals. Though our definition of altruism (p. 110) refers only to ingroup interactions, in our model individuals interact with outgroup individuals as well: the model works because altruists confer fitness advantages on insiders, while inflicting fitness costs on outsiders. Our references to “group beneficial” behaviors thus refer exclusively to in-group effects.

We thus address two questions: what accounts for the evolution of individually costly and in-group-beneficial forms of human sociality towards non-kin? And what accounts for the differential success of those common group-level institutional structures such as states, resource sharing, and monogamy which have emerged and proliferated repeatedly and in a wide variety of circumstances during the course of human history? The co-evolutionary process that we model and simulate are based on the idea that the two questions may be more convincingly resolved jointly than singly.

8.2 Selective Extinction

Consider a large meta-population of individuals living in partially isolated sub-populations (called demes). As in Figure 3.1, we assume that altruists (A’s) take an action costing the individual c that confers a benefit b on an individual randomly selected from the n members of the deme. The model and notation are summarized in Table 8.2. A’s are bearers of a hypothetical ‘altruistic allele’; those without the allele (N’s) do not behave altruistically. Reproduction is asexual. In the absence of reproductive leveling, individual fitness is identical to the prisoner’s dilemma payoffs in Figure 3.1. For example, an A in a group with all A's will on average receive benefits b and hence will expect a number of offspring surviving to reproductive age, that is $b - c$ greater than the fitness of an N who interacts only with N’s.
(1) Generic Price Equation:
\[ \Delta p = \text{var}(p_j) \beta_G + \text{var}(p_{ij}) \beta_i \]
(2) Generic condition for \( A \) to increase:
\[ \frac{\text{var}(p_j)}{\text{var}(p_{ij})} = \frac{F_{ST}}{(1 - F_{ST})} > -\frac{\beta_i}{\beta_G} \]
(3) Effect of \( A \) on deme-average fitness
\[ \beta_G = \frac{d w_j}{dp_j} = \kappa \left( \frac{d w_j}{d \lambda} \right) \left( \frac{d \lambda}{dp_j} \right) = 2 \kappa \lambda_A \]
(4) Effect of \( A \) on individual fitness:
\[ \beta_i = \frac{d w_{ij}}{dp_j} = -(1 - \tau)c + \frac{b}{n} + \frac{2 \kappa \lambda_A}{n} \]
(5) Price’s Equation for the model:
\[ \Delta p = \text{var}(p_j) 2 \kappa \lambda_A - \{(1 - \tau)c - (b + 2 \kappa \lambda_A) \}/n \]
(6) Condition for \( A \)'s to increase (\( n = \infty \)):
\[ \frac{F_{ST}}{(1 - F_{ST})} > (1 - \tau)c/2 \kappa \lambda_A \]

**Table 8.1. Summary of model and notation.** \( b \) and \( c \) are benefits and costs to deme members; \( p_k \) = percent of deme \( k \) that are \( A \)'s; and \( p = \) percent of metapopulation that are \( A \)'s. In equation (1), \( \Delta p \) = between-deme effect + within-deme effect; \( F_{ST} \) = between-deme variance/total variance; \( \kappa \) = probability of interdemic contest; \( \tau \) = extent of reproductive leveling; Larger \( F_{ST} \) favors \( A \)'s. Equation 6 says that \( F_{ST}/(1 - F_{ST}) > \) individual cost/deme benefits.

We want to determine the conditions under which \( p \), the fraction of \( A \)s in the meta-population, will increase. Recall from equation (6.1) that whether altruism evolves (\( \Delta p > 0 \)) depends on the outcome of a race in which the between-deme selection process promoting its spread, namely \( \text{var}(p_j) \beta_G \), competes with the within-group selection process tending to eliminate it, namely \( \text{var}(p_{ij}) \beta_i \). For the between-deme effect to exceed the within-deme effect (rearranging equation 6.1), it must be that:

\[ \frac{\text{var}(p_j)}{\text{var}(p_{ij})} > -\frac{\beta_i}{\beta_G} \]

The right hand side of 8.1 is the ratio of two fitness effects: \( \beta_i = dw_{ij}/dp_{ij} \), the effect on \( i \)'s fitness of switching from \( N \) to \( A \), and \( \beta_G = dw_j/dp_j \), the effect on average fitness of members of group \( j \) on variations in the fraction group \( j \) who are \( A \)'s. A behavior is altruistic if adopting it lowers one’s expected fitness while increasing the average fitness of one’s
deme (§A8). Given this definition, we are interested in the case where \( \beta_i \) is negative and \( \beta_G \) is positive. The left hand side of (8.1) is a measure of positive assortment arising from the fact that if the fraction of A’s in demes differ (that is, \( \text{var}(p_j) \) is positive), then A’s are more likely than N’s to interact with A’s. Because the within-deme benefits of altruism are randomly distributed, between-deme differences in the prevalence of A’s (i.e. \( \text{var}(p_j) > 0 \)) is the only reason why A’s are more likely than N’s to interact with A’s, and thus to benefit mutually.

But, if A’s are likely to benefit for this reason, they are also more likely to compete over deme-specific resources (Taylor 1992, Wilson et al. 1992). In the selective emigration model of Chapter 6 we allowed the predominantly altruistic demes to grow without these resource constraints. But, here we assume the most stringent form of local density-dependent constraints on reproductive output: sites are saturated so that territorial expansion—gaining a new site at some other group’s expense—is required for deme growth. Thus altruism can proliferate only by helping a deme to acquire more territory, not by any of the other ways that members of predominantly altruistic demes might produce more surviving offspring.

Selective extinction may allow the evolution of altruism if predominantly altruistic demes are more likely than other demes to survive between-deme contests and to colonize and repopulate the sites vacated by demes that fail (Aoki 1982). This process is captured by the term \( \beta_G \), the size of which is the product of the determinants: the frequency of contests, the fitness effects of surviving a contest, and the contribution of altruists to surviving. In every generation with probability \( \kappa \) each deme engages in a ‘contest.’ A contest may be a hostile encounter or an environmental challenge without direct deme interaction. Demes that fail are eliminated and surviving demes repopulate the vacated sites.

Demes are the same size (normalized to 1) except that demes that have occupied the site of an eliminated deme are momentarily of size 2 (and eliminated demes are of size zero). The surviving deme then divides, forming two daughter demes of equal size. Let the probability that the deme survives be \( \lambda \). The size of deme \( j \) next generation is thus 1, 2 or 0 with probabilities \( 1 - \kappa, \kappa \lambda \) and \( \kappa (1 - \lambda) \), respectively, so expected size is \( w_j = 1 - \kappa + 2 \kappa \lambda \). The effect of the prevalence of A’s on the expected size of the deme next generation; that is, \( \beta_G = dw_j/dp_j \) is the likelihood of a contest (\( \kappa \)), times the effect on deme size of surviving or not (2), times the effect of the prevalence of A’s on the probability of a deme surviving should a contest occur.
(\(\lambda_A\)); so \(\beta_G = 2\kappa\lambda_A\). There is no way to estimate \(\lambda_A\) empirically, so we explore two alternative values: \(\lambda_A = 1\) is derived from a model in which all-\(A\) and all-\(N\) demes (respectively) survive and fail with certainty should a contest occur; while if \(\lambda_A = 1/2\), an all-\(A\) deme survives with probability 3/4 and an all-\(N\) deme survives with probability 1/4. These two alternatives are shown in Figure 8.1.

![Figure 8.1](image_url)

Figure 8.1. *Survival probability for deme \(j\) if half of the members of the paired group are \(A\)'s.*

### 8.3 Reproductive Leveling

Individual differences in size, health, behavior, and other influences on access to scarce resources are typically reflected in differences in reproductive success. Among some other primates (Noe and Sluijter 1995, Pandit and van Schaik 2003) and especially among humans, reproductive leveling attenuates this relationship. Because altruists in the model receive lower payoffs than other deme members, they benefit from reproductive leveling because this attenuates the within-group selective pressures working against them, reducing \(\beta_i\).

To see how this works, suppose a particular \(N\) were instead an \(A\). In the absence of reproductive leveling and intergroup competition its fitness would be less by an amount \(c\), the cost of the altruistic behavior. But the individual who switched from \(N\) to \(A\) would also have a \(1/n\) chance of interacting with itself, garnering the randomly assigned benefit \(b\). Additionally, by switching from \(N\) to \(A\), the individuals contributes directly to its fitness in another way: by increasing the chance of survival of the deme, in which case, like every member of the surviving deme, it will be doubled. The
additional direct effect of the switch from $N$ to $A$ is the resulting change in $p_j$, namely $1/n$, times the effect of variations in $p_j$ on deme-average fitness $\beta_G$. Thus

$$\beta_i = \frac{dw_{ij}}{dp_{ij}} = -c + \frac{b}{n} + \frac{2\kappa \lambda_A}{n},$$  \hspace{1cm} (8.2)$$

Recall that reproductive leveling is a convention, adherence to which is in the interest of each deme member as long as most others follow the convention. Reproductive leveling takes the form of a proportional deduction at rate $\tau$ of each members’ payoffs, the proceeds of which are distributed equally to all members. An example would be ‘common pot’ food sharing in which some fraction of the food available to a group is pooled irrespective of who acquired it, and then distributed in equal shares to each member. The effect is to reduce within-deme fitness differences between $A$’s and $N$’s from $-c$ to $-(1-\tau)c$, so taking account of all of the direct effects of a switch from $N$ to $A$ on the actors fitness we have.

$$\beta_i = -(1-\tau)c + \frac{b}{n} + \frac{2\kappa \lambda_A}{n}.$$  \hspace{1cm} (8.3)$$

In keeping with our representing the $A$’s as altruistic, we choose parameters such that $\beta_i$ is negative. Figure 8.2 reproduces the information in Figure 6.1 but suppresses the baseline fitness, and shows the effect of resource sharing on the payoff differences of the two types. The difference in the probability of meeting an $A$, conditional on one’s own type, that equalizes expected payoffs is no longer $P(A|A) - P(A|N) = F^*$ as shown by the solid lines in Figure 8.4, but is now given by the dashed lines and is $P(A|A) - P(A|N)' = F' < F^*$. Comparing the two figures one sees that $F^* = c/b$ while $F' = c(1-\tau)/b$. As a result, were the population structure as in Figure 8.2, the degree of genetic differentiation given by $(F^*)$, and the sharing institution in place ($\tau > 0$), then $\omega^A > \omega^N$, so $p$ would increase.

Using these values for $\beta_i$ and $\beta_G$ into the Price equation 6.1, and using the definition of $F_{st}$ as the fraction of the total genetic variance that is between groups, we rewrite equation 8.1 and find that the $A$’s share of the meta-population will increase if

$$\frac{F_{st}}{1-F_{st}} > \frac{\beta_i}{\beta_G} = \frac{(1-\tau)c - b/n}{2\kappa \lambda_A} - \frac{1}{n}.$$  \hspace{1cm} (8.4)$$
Figure 8.2. Resource Sharing Reduces the Level of Positive Assortment Necessary for the Evolution of an Altruistic Trait. Note: The horizontal axis measures the probability of being paired with an A. The dashed payoff functions indicate the effect of within-group resource sharing. For the value of τ illustrated here, the altruistic trait will proliferate if $F > P(A|A') - P(A|N')$ while for $τ = 0$, $F > P(A|A) - P(A|N)$ is required. For the case of nonrandom pairing within demes (to be introduced below in the simulations reported in §8.7), the analysis is similar, with $b - cτ$ on the right vertical axis replaced by $(1 - ζ)b$ and $-(1 - τ)c$ on the left axis replaced by $ζb - c$.

If $n$ is large, this is approximated by the much simpler

$$\frac{F_{ST}}{1 - F_{ST}} > \frac{(1 - τ)c}{2κλ_A}.$$  \hfill (8.5)

Like Hamilton’s rule for the evolution of altruism by inclusive fitness, this model thus yields a condition indicating the minimum degree of positive assortment necessary to allow altruism to proliferate. The left hand term is analogous to Hamilton’s $r$, the degree of relatedness, but here positive assortment arises solely from between-deme differences in the prevalence of $A$’s, not from the tendency of recent descendant kin to interact preferentially. The right hand terms in equations 8.4 and 8.5 are the ratio of individual costs to group-level benefits (analogous to $c$ and $b$ respectively in Hamilton’s rule).
We now ask if ancestral humans are likely to have lived under conditions such that equations 8.4 or 8.5 would be satisfied. Table 8.2 is a summary of the main parameters and the estimated range of empirically plausible values.

<table>
<thead>
<tr>
<th>Determinant</th>
<th>Range explored</th>
<th>Comment or Method of Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-demic genetic differentiation</td>
<td>$F_{ST}$</td>
<td>Genetic markers (recent foragers)</td>
</tr>
<tr>
<td>Reproductive Leveling</td>
<td>$\tau$</td>
<td>Food sharing (recent foragers)</td>
</tr>
<tr>
<td>Gains-losses from contests per generation</td>
<td>$\omega$</td>
<td>Archeological and ethnographic evidence</td>
</tr>
<tr>
<td>Per generation probability of a decisive (2,0) contest</td>
<td>$\kappa$</td>
<td>Based on Estimates of mortality in on-going conflict</td>
</tr>
<tr>
<td>Effect of percent altruists on deme survival</td>
<td>$\lambda_A$</td>
<td>Arbitrary (see Figure 7.1)</td>
</tr>
<tr>
<td>Effective deme size (one generation)</td>
<td>$n$</td>
<td>Coalition of bands</td>
</tr>
<tr>
<td>Cost to Altruist</td>
<td>$c$</td>
<td>Depends on behavior under consideration</td>
</tr>
<tr>
<td>Benefits to deme (w/o a contest)</td>
<td>$b$</td>
<td>As immediately above (virtually no effect)</td>
</tr>
</tbody>
</table>

Table 8.2. **Parameter estimates.** Benchmark values are in bold. Entries not in bold are alternative values ($\omega = 0.4$ not used).

### 8.4 Empirical estimates of $F_{ST}$

Wright (1950):203 speculated that an equilibrium $F_{ST}$ among human groups, namely that which would balance the offsetting effects of migration and genetic drift, might be about 0.02, a value that would preclude inter-demic competition as an important evolutionary force. For example, in the selective emigration model an $F_{ST}$ of this magnitude would require $b$ to exceed $c$ by a factor of 50 in order for altruism to evolve (equation 6.9). But, as we
have seen in the previous chapter, most empirical estimates of $F_{ST}$ values are considerably larger.

However, estimates of genetic differentiation at the locus of an allele that is expressed in an altruistic behavior may differ from estimates based on neutral loci (those not under selection) such as those in Table 7.1. First, an altruistic allele would be, by definition, under directional selection. This would be expected to reduce inter-demic genetic differentiation at least in the very long run, because in the absence of offsetting effects, the frequency of the $A$’s in the population will eventually go to zero. However this tendency may not work over time scales relevant to human demes. Simulations show that even for very strong selection against the $A$’s and for plausible initial distributions of $A$’s in demes, the $F_{ST}$ rises for tens of generations (Bowles 2006). For moderate selection against the $A$’s, the $F_{ST}$ may rise for more than a hundred generations before falling. Because, as we found in the previous chapter, fission and extinction events that enhance inter-demic variance are likely to be an order of magnitude more frequent than this, it appears that high levels of $F_{ST}$ could persist indefinitely. Additional simulations show that exceptionally strong directional selection against the $A$’s ($c = 0.1$) is compatible with the indefinite maintenance of high levels of $F_{ST}$, even with random rather than “associative” fission of demes.

Second altruists will sometimes be able exclude non-altruists’ from their demes (Wilson and Dugatkin 1997), resulting in what Eshel and Cavalli-Sforza (1982) called ‘selective assortment.’ In this case, migration might also enhance between-deme variance and reduce within-deme variance (Rogers and Jorde 1987). Here, selective assortment is contingent on past behavior that is itself an observable expression of the altruistic allele. As a result, the only way an $N$ could mimic the $A$’s so as to evade their choosiness would be to adopt the altruistic behavior itself and thus to bear its costs. Thus the instability arising in the case of assortment by ‘green beards’ (Ridley and Grafen 1981) does not arise.

But, there is nonetheless an impediment to selective assortment that is sometimes overlooked: exclusion of $N$’s is likely to be costly for the $A$’s, while the associated benefits are shared by all deme members. However, it is not implausible that altruists would undertake some moderate level of $N$-exclusion as a contribution to the public good. As we have seen, there is ample ethnographic evidence that foragers practice selective assortment when they ostracize or shun individuals who violate behavioral norms. Finally,
it is readily shown that a modest amount of selective assortment generates substantial levels of between-deme differences (Bowles 2006).

8.5 Deme Extinction and the Evolution of Altruism

While the effects of most forms of reproductive leveling cannot be estimated, the degree of within-deme resource sharing is known from empirical studies of the acquisition and consumption of nutrition among foragers. These and the other empirical data mentioned here are summarized in Bowles (2006). On this basis we take $\tau = 2/3$ as a plausible benchmark with one-third an alternative value. The appropriate value of $n$ is the number of deme members of a breeding generation (about a third of the census size). Individual bands may have competed for survival, but it is likely that bands in coalition also engaged in contests. Here we take 96 members as a benchmark size, or $n = 32$ members of a single generation. As an alternative, we will also consider very large (strictly, infinite) demes. Plausible values of $c$ and $b$ will depend on the particular altruistic behavior in question. For example, a warning call would have a different $b$ and $c$ than defending the community against hostile neighbors. To facilitate the exploration of a variety of altruistic behaviors, we present results for a given $b = 0.05$ and $c$ varying from 0 to 0.08. Equations 8.4 and 8.5 make it clear that for sizable demes $b$ is of little importance. The main contribution of altruistic behavior to fitness arises from group competition, not within-group benefits.

The extent of hostile group interactions during the late Pleistocene and early Holocene surveyed in the previous chapter show that the level of ongoing hostility may have resulted in the equivalent of a contest in which the loser is eliminated occurring every five or seven generations. Here we use the lower estimate.

The above parameter estimates are summarized in Table 8.2. They are subject to substantial error given that they are inferences about conditions occurring tens of thousands of years ago for which very little direct evidence is available.

With this caveat in mind, suppose early human demographics and social practices resulted in genetic differentiation at the locus of an altruistic allele that was the magnitude of the median in Table 7.1 (0.076). For the benchmark values of $\tau$, $n$, and $\lambda_A$, the solid lines in Figure 8.5 give the combinations of $c$ and $\kappa$ such that equations 8.4 and 8.5 are satisfied as
equalities. More frequent contests or less costly forms of altruism (points above the line) allow altruism to proliferate. Dashed lines do the same for more stringent alternative parameter values. For example, for the estimated $\kappa$, if $c = 0.05$, altruism proliferates under both survival functions and the benchmark assumptions, but not for very large demes with limited reproductive leveling.

8.6 The Australian Laboratory

To study the evolutionary consequences of warfare under Pleistocene conditions using recent data, one would ideally use estimates of both genetic differentiation and wartime mortality from hunter-gatherer populations living in close proximity with one another but having little contact with farmers or herdsmen. Such groups exist in Arnhem Land, Australia. Australia is the continent thought by many to be the best laboratory of likely Pleistocene and early Holocene conditions among foragers (Lourandos 1997). Depictions of warriors and battles in the rock art of Arnhem Land populations date from as early as ten thousand years ago (Tacon and Chippendale 1994). The availability of archaeological, ethnographic and genetic data for this region makes it a remarkable laboratory for this investigation. Table 8.3 presents data on the extent of wartime mortality in three nearby groups of foragers, the Anbara, Murngin, and Tiwi, along with estimates of genetic differentiation among seven Aboriginal groups, including the Tiwi and the Murngin, in that relatively small area, depicted in Figure 8.3.

To gauge their implications we rearrange equation 8.4 and for simplicity set $b = 0$ to define the critical value $c^*$ as the maximum cost of the altruistic behavior consistent with its proliferating in the population:

$$c^* = 2\kappa \lambda_A \left( \frac{F_{ST}}{1 - F_{ST}} + \frac{1}{n} \right). \quad (8.6)$$

For these populations, we use an estimated value of $F_{ST} = 0.040$, estimated for groups in this region, that is considerably lower than the mean of those for hunter-gather groups generally presented in Table 7.1. The estimates of $c^*$ for these populations, assuming two values of $\lambda_A$, make it clear that if groups were as differentiated as these populations and as warlike as the Murngin, between-group competition could overcome very strong within-group selection against an altruistic behavior. Even for groups
similar to the more peaceful Anbara, quite costly forms of altruism could proliferate by this mechanism ($c^* = 0.015$ for $\lambda_A = 1$).

<table>
<thead>
<tr>
<th></th>
<th>Murgnin</th>
<th>Tiwi</th>
<th>Anbara</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta$</td>
<td>0.207</td>
<td>0.100</td>
<td>0.0045</td>
</tr>
<tr>
<td>$\lambda_A$</td>
<td>1</td>
<td>0.066</td>
<td>0.032</td>
</tr>
<tr>
<td>$\lambda_A$</td>
<td>0.5</td>
<td>0.033</td>
<td>0.016</td>
</tr>
</tbody>
</table>

Table 8.3. **When an Altruistic Trait Can Proliferate.** The largest cost $c^*$ for an altruistic trait to proliferate given genetic differentiation and mortality in intergroup hostilities ($\delta$) among three Arnhem Land Australian hunter gatherer populations. Note: We assume here that $\kappa = 2\delta$. Genetic differentiation ($F_{ST} = 0.040$) is among seven groups, including Tiwi and Murgnin. See Bowles (2009a); $\delta$ is from Table 8.3; $n = 26$ effective size ($N_e$) based on the typical size calculations for non-equestrian, non-Arctic foragers during the Late Pleistocene. See the previous chapter and Marlowe (2005).
8.7 The Coevolution of Institutions and Altruism

The data in Figure 8.4 and Table 8.3 make it clear that for many populations and for plausible parameter values, genetic differentiation is such that even very infrequent contests would have been sufficient to spread quite costly forms of altruism. Because the initial spread of altruism among humans could have been propelled by just a few of the vast number of late Pleistocene demes, the above data and reasoning suggest that selective deme extinction may be part of the account of the evolution of altruism. This is likely in the presence of significant levels of reproductive leveling, but not in its absence, suggesting an important role for culturally transmitted practices in creating a niche in which a genetic predisposition to behave altruistically might have evolved, and perhaps accounting for the distinctive aspects of human altruism not found in other species.

But, we have not yet modeled the dynamic of genes and group level institutions. Thus far we have posited empirically plausible levels of warfare and reproductive leveling, rather than explained their evolution. Below and in the next chapter, we will model the coevolution of altruism with these distinctive human institutions. Because this process involves highly complex selection processes operating at two levels, individual and group, in which the magnitude of within-group and between-group selection effects are endogenously determined by the evolution of group-level institutions, the dynamic is not amenable to mathematical analysis using the Price equation, or indeed any other mathematical formulation capable of analytical solution. An agent-based model (§A1), however, is illuminating.

Our objective is to see if the culturally transmitted group-level institutions that support the evolution of the ‘altruistic allele’ could themselves have evolved even if maintaining these institutions imposed costs on the groups adopting them. We will of course consider reproductive leveling. But, we first introduce an additional aspect of group social structure, the tendency of like types to interact within groups more frequently than would occur by chance. The resulting non-random pairing, sometimes called segmentation, may occur because of a tendency to interact with siblings and other kin by recent common decent, or because people can sometimes condition their interactions on the type of the other, based on information about past behavior, for example.
Figure 8.4. **Conditions for the evolution of altruism by selective extinction and reproductive leveling.** The solid lines are the benchmark values estimated in the text \( n = 32, \tau = 0.66 \). Line 1: \( n = 4, \tau = 0.33 \); Line 2: \( n = 32, \tau = 0.33 \). Line 3: \( n = 4, \tau = 0.66 \). Points above the each line give combinations of \( c \) and \( \kappa \) such that altruism would proliferate according to equations 8.4 and 8.5. Panels a and b (respectively) use \( \lambda_{\text{A}} = 0.5 \) and \( \lambda_{\text{A}} = 1 \). For both panels, \( b = 0.05 \). Similar analysis for all of the data in Table 7.1 is presented in Bowles (2006).

Suppose, then, that in addition to the institution of resource sharing, groups are also segmented, so that rather than the benefit of a given altruist’s helping behavior being bestowed upon by an individual randomly drawn from the group we now let altruists have a higher than by chance probability of being the recipient of the help provided by a particular altruist, rather than being paired with an N and receiving nothing. Let \( \zeta > 0 \) be the degree of segmentation, that is, the difference in the conditional probability of an \( A \) being paired with an \( A \) and an \( N \) being paired with an \( A \) in the within-group pairing. Thus, the probability of an \( A \) in group \( j \) being paired with an \( A \) is \( \zeta_j + (1 - \zeta_j)p_j > p_j \) and the probability of an \( N \) being paired with an \( A \) is \( (1 - \zeta_j)p_j < p_j \). So the expected benefit of the presence of altruists in the group enjoyed by an altruist is no longer \( p_j \hat{b} \) but
rather \(\{\xi + (1 - \xi) p_j\}b\), and the expected benefit of the non-altruist is only \((1 - \xi) p_j b\). Then, abstracting from reproductive leveling, the difference in payoffs between \(A\)'s and \(N\)'s in the same group is no longer \(c\), but \(c - \xi_j b\). Segmentation thus reduces the fitness disadvantage of altruists because they are disproportionately likely to benefit from the help of other altruists, while \(N\)'s are disproportionately likely to meet other \(N\)'s. If \(\xi > c/b\), \(A\)'s will on average do better than \(N\)'s in group \(j\) and as a result the \(A\)'s will not suffer any reproductive disadvantage. To pose the classical (strong) group-selection problem, we therefore assume \(\xi < c/b\). Like resource sharing, segmentation is a convention and is passed on culturally.

Taking account of both segmentation and resource sharing, the differences in the expected payoffs received by \(N\)'s and \(A\)'s within a group will now be \((1 - \tau)(\xi b - c)\) from which it is clear that both institutions retard the within-group selection against the \(A\)'s.

The institutions represented by \(\xi\) and \(\tau\) differ among groups and they also evolve under the selective pressure resulting from the differential survival of groups. In our simulations, when conflict occurs between groups, the group with the higher total payoff wins. The losing group’s members die and the winning group populates the site occupied by the losers with replicas of a random selection of themselves. The new inhabitants of the site adopt the institutions of the winning group from which they descended. Institutions are also subject to stochastic variation, increasing or lowering \(\tau\) and \(\xi\) by chance each period. Both segmentation and resource sharing impose costs on the groups adopting them. More segmented groups may fail to capture the benefits of diversity or of economies of scale, and resource sharing may reduce incentives to acquire the resources to be shared. Neither of these costs are modeled formally, but to capture their impact, group average benefits are reduced by an amount that is rising in convex in both \(\xi\) and \(\tau\).

With Astrid Hopfensitz and Jung Kyoo Choi, we simulated an artificial population living in 20 groups (Bowles et al. 2003). We selected an extraordinarily high cost of altruism, \(c = 0.1\) and did not allow \(A\)'s to benefit from their own prosocial actions; that is, \(b = 0.2\) is conferred on some other randomly selected group member. Thus within a group, when \(\xi = \tau = 0\), \(A\)'s have ten percent fewer offspring than \(N\)'s, making these simulations a demanding test of our interpretation. These and the other benchmark values of the parameters in the simulations shown in Table 8.4 were chosen on grounds of empirical plausibility. We initiated each simulation with altruists
and institutions absent at time zero, to see if they would proliferate if initially rare (the individual and institutional mutation process will introduce some variability in the population).

<table>
<thead>
<tr>
<th></th>
<th>Benchmark Values</th>
<th>Range Explored</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Group Size ($n/g$)</td>
<td>20</td>
<td>7 to 47</td>
</tr>
<tr>
<td>Migration Rate ($m$)</td>
<td>0.2</td>
<td>0.1 to 0.3</td>
</tr>
<tr>
<td>Probability of Conflict ($k$)</td>
<td>0.25</td>
<td>0.18 to 0.4</td>
</tr>
<tr>
<td>Mutation Rate ($e$)</td>
<td>0.001</td>
<td>0.1 to 0.000001</td>
</tr>
</tbody>
</table>

Table 8.4. Key Parameters for the Simulation. Note: Total population size is $n$, and there are $g$ groups; $m$, $k$, and $e$ are per generation. Other parameters: benefit $b = 2$, cost $c = 1$; baseline payoffs $= 10$. We varied group size by varying $n$. For reasons explained in the text, we restricted $s$ to not exceed 0.5 while $\tau \in [0, 1]$. The costs imposed on the group by these institutions are $(\xi^2 + \tau^2)/2$.

### 8.8 Simulating Gene-culture Coevolution

A simulation of this model appears in Figure 8.5. The early rise in $p$ is supported by the chance increase in both $\xi$ and $\tau$ (between periods 100 and 150). When $p$, the fraction of the meta-population who are $A$’s, reaches high levels (periods 532 to 588, for example) both $\xi$ and $\tau$ decline, typically leading to a sharp decline in $p$. The subsequent rise in $\xi$ or $\tau$ occurs by chance. The pattern emerges for the following reason. When the population is evenly divided between $A$’s and $N$’s, many groups are also approximately evenly divided, which means that $\overline{\text{var}}(p_{ij})$ takes its maximum value, so the within-group selection term in the Price equation (6.2) is also at its greatest. As a result, the beneficial effects of retarded within-group selection gained by higher levels of $\tau$ or $\xi$ are maximized in this region. By like reasoning, when $p$ is well above 0.5, the benefits of the protection of $A$’s offered by the institutions is of less value. But the institutions are costly to bear so when $p$ is high, groups with substantial levels of segmentation or resource sharing are likely to lose conflicts with other groups, and the sites they had occupied are then peopled by the descendants of winners, who typically bear lower levels of these institutional variables. As a result, both $\xi$ and $\tau$ fall.

We checked to see if the proliferation of the $A$’s could have been the result of the direct fitness benefits to an individual of switching from $N$ to
The Coevolution of Leveling, Within-group Segmentation, and Altruism. The figure represents 1000 period history of a run using the benchmark parameters from Table 8.2. The population average frequency of altruists is $p$, while $\tau$ and $\zeta$ give the average across the 20 groups of the level of resource sharing and segmentation. Altruism and both group-level institutions are initially rare. The particular time frame shown was selected because it clearly reveals this dynamic, which is observed over long periods in many runs.

Could individual-level altruism have evolved had group level institutions not co-evolved with it. To answer this question we constrained $\zeta$ and $\tau$ to be zero in all periods. Altruism failed to evolve over a large number of implementations of this treatment. We also investigated whether the.
institutions would evolve if \( p \), the fraction of altruists in the population, is constrained to zero. They do not, because institutions are costly and where there are no altruists in the population the institutions perform no group-beneficial function, thus leading groups that by chance adopt a high level of sharing or segmentation to lose conflicts in which they are involved.

Finally, we wanted to know how sensitive our simulations are to variations in the key parameters. To do this, we varied group size from 7 to 47, and for each size ran 10 simulations of 50,000 generations, with the other parameters at their baseline values. We did this with both institutions constrained to not evolve, with each singly constrained to not evolve, and with neither constrained. We performed the same operation for variations in the migration rate from 0.1 to 0.3, and the probability of conflict (\( \kappa \)) from 0.18 to 0.51. The results appear in Figure 8.6.

The top panel shows that with both institutions constrained not to evolve, a group size of 7 (meaning about 21 individuals, roughly the size of a small foraging band) supports high levels of altruism, but group sizes greater than 8 result in a frequency of altruists of less than 0.3. This is a highly improbable scenario under which altruism could evolve as a group of this size would surely have a migration rate greater than our benchmark of 0.2. Taking as a benchmark the group size for which \( p > 0.5 \), we see that with no institutions the critical size is 8, while with both institutions \( p > 0.5 \) for all group sizes less than 22. The results for the migration rate are similar. Sustaining \( p > 0.5 \) requires a (per generation) migration rate of not greater than 0.13, but with both institutions free to evolve, the critical migration rate is 0.21. The bottom panel shows that institutions also allow the evolution of high levels of altruism with significantly fewer between-group conflicts. A “vertical” reading of the figure is also illuminating: for example, the bottom panel shows that for \( \kappa = 0.3 \), \( p \) is less than 0.2 without institutions, but is greater than 0.8 with both institutions free to evolve.
Figure 8.6. **Group-level Institutions Increase the Size of the Parameter Space for which Altruistic Behaviors are Common.** Each data point is the population average frequency of altruists over 10 runs of 50,000 periods each for the parameter value indicated on the horizontal axis. Each run began with $p = \tau = \zeta = 0$. The curve labeled “none” gives the results for runs in which $\tau$ and $\zeta$ were constrained to zero; the other curves indicate runs in which one or both of the institutions were free to evolve. (“Tax” refers to resource sharing.) The horizontal distance between the curves indicates the enlargement of the parameter space made possible by group level institutions. The vertical distance between the curves shows the impact of institutions on $p$. 
8.9 Levelling, Competition, and Cooperation

We have shown that the between-group genetic differentiation, patterns of group conflict and other conditions under which our model of selective extinction with reproductive leveling could have worked to proliferate altruism are very likely to have existed at least for some early human groups. We have also described a process whereby institutions such as reproductive leveling and within-group segmentation provide an environment within which an individually costly group-beneficial trait may evolve, and in which these institutions proliferate in the population because of their contribution to the evolutionary success of the group-beneficial trait.

Our simulations have shown that if group-level institutions implementing resource sharing or positive assortation within groups are free to evolve, group-level selection processes support the co-evolution of group-beneficial individual behaviors along with these institutions, even where these institutions impose significant costs on the groups adopting them. In the absence of these group-level institutions, however, group-selection pressures support the evolution of group-beneficial traits only when intergroup conflicts are very frequent, groups are small, and migration rates are low. Our simulations also suggest that both altruistic behaviors and reproductive leveling institutions could have emerged and proliferated from an environment in which they were initially rare. Recall that \( p, \tau, \) and \( \zeta \) are all set to zero at the beginning of our simulations and all three evolve from this initial condition, as in Figure 8.6.

Crucial to this process is the fact that food sharing and other forms of reproductive leveling do not require the pre-existence of altruistic preferences: adherence to these conventions is a best response for self-interested individuals. Thus, as we have seen in Chapter 3, a norm that prohibits the powerful from monopolizing food supplies could be sustained by the repeated nature of dyadic network interactions or the adverse reputations that violators would suffer. In light of the significant benefits of co-insurance in the highly variable Pleistocene environments, the conditions outlined in Chapter 3 for the success of reciprocal altruism or indirect reciprocity—frequent ongoing interactions with public and high quality information—could easily been met, allowing the initiation of food sharing, especially in relatively small networks of closely related individuals. Indeed, as we have seen in Chapter 1, it seems likely that some kind of food sharing is a very early development in human behavior, having co-evolved with the
shift to diet based on a large meat packages. Once established the norm could persist in larger, less closely related groups, as Axelrod and Hamilton (1981) showed. Its effect, as this chapter has shown, would be to promote the spread of altruistic preferences including the public spirited punishment of transgressors, thereby stabilizing reproductive leveling even for groups of more substantial size than we have considered here.

Notwithstanding the highly speculative nature of these inferences, it seems possible that the social and physical environments of the late Pleistocene may fall within the parameter space supporting the co-evolutionary trajectories illustrated in Figure 8.6. If so, the multi-level selection model with endogenous institutions may provide at least a partial account of the evolution of individually altruistic behaviors as well as group level resource sharing, segmentation and perhaps other institutions during this critical period of human evolution.

The main causal mechanisms of the model—institutionalized resource sharing among non-kin and intergroup conflict—suggest a central role for uniquely human cognitive, linguistic and other capacities in this process, perhaps helping to account for the distinctive levels of cooperation among non-kin practiced by humans. The same observation suggests the limited applicability of the model and simulations to most other animals. However, as we have seen, reproductive leveling occurs in some other species, other forms of within-group variance reduction are also not uniquely human. Moreover, for species in which neighboring groups including unrelated members compete for resources or in which group extinctions are common, a similar model might apply. In these cases individually costly group-beneficial behaviors may contribute by increasing group size, an effect we have not considered here, or in other ways, to the success of the group in avoiding extinctions or in gaining resources from neighboring groups.

Examples include social mammals such as the cooperative mongoose *Suricata suricatta*, for which group extinction rates are inversely correlated with group size and in some years exceed half the groups under observation (Clutton-Brock, Gaynor, McIlrath, Maccoll, Kansky, Chadwick, Manser, Skinner, and Brotherton, 1999). Similarly, fire ants (*Solenopsis invicta*) and a large number of other ant species form breeding groups with multiple unrelated queens and practice brood raiding and other forms of hostility toward neighboring groups, with success positively related to group size (Bernasconi and Strassmann 1999). Mortality in intergroup conflict among
chimpanzees may as great as our estimates for humans, with more powerful groups gaining fitness-enhancing territories (Nishida et al. 1985, Manson and Wrangham 1991, Mitand et al. 2010). Whether the levels of cooperation observed in these and other species—cooperative breeding in meerkats or the within-group cooperation of only moderately related male chimps, for example—might be explained in part by the causal mechanisms at work in our model is an interesting question which we have not explored.

Our simulations show that two aspects of the constructed niches that supported the evolution of altruistic individuals, reproductive leveling and within-group segmentation, could have coevolved with altruism. But lethal group conflict also plays a central role in this explanation, and thus far we have not sought to explain either the motives promoting it or the structure of group interactions in which lethal conflict is a likely outcome. This is our next task.
9

Parochialism Altruism and War

Who trusted God was love indeed
And love Creation’s final law
Tho’ Nature, red in tooth and claw
With ravine, shriek’d against his creed.

Alfred, Lord Tennyson,
_In memoriam, A.H.H._
_Canto 56_ (1849).

For Alfred, Lord Tennyson, love and religious fealty were man’s triumph over a violent and recalcitrant Nature. But late nineteenth century scientists as diverse as Charles Darwin (1873) and Karl Pearson (1894) recognized war as a powerful evolutionary force that paradoxically might account for social solidarity among humans and altruism towards the fellow members of one’s group. The previous chapter confirmed that inter-group conflict may contribute to the evolution of altruism.

Did Lord Tennyson get it wrong?

Was love the claw’s unwonted child?
Creation, too, the spawn of strife?
That creed sustained enduring life
So let not Nature be defiled.

The idea that altruism would spread by means of war has not been subjected to systematic investigation. When inter-group conflict has been considered, its extent has typically been assumed, as in the previous chapter, rather than explained. Recent contributions have shown that insider favoritism could evolve if it facilitates generalized exchange (Yamagishi et al. 1999), or supports the higher payoffs that occur when people with similar norms interact (McElreath et al. 2003), or coordinates the efficient selection of particular ways of interacting (Axtell et al. 2001), or improves communication among group members so as to facilitate informal enforcement of
prosocial norms (Bowles and Gintis 2004). Related evolutionary explanations of why group boundaries so powerfully influence human behavior are found in Nettle and Dunbar (1997) and Hammond and Axelrod (2006). But the evolution of hostility towards outsiders and other conditions that might account for the distinctly lethal and common nature of human warfare has received little attention.

The importance of insider favoritism in human behavior is well established in experiments, as we have seen in Chapter 5. Other evidence demonstrates that individuals often favor fellow group members over ‘outsiders’ in the choice of friends, exchange partners and other associates and in the allocation of valued resources (Brewer and Kramer 1986, McPherson et al. 2001).

We take as emblematic of this evidence a ‘third party punishment’ experiment with subjects drawn from two nearby, but not recently hostile, linguistic groups in the famously bellicose highlands of Papua New Guinea (Bernhard et al. 2006). Recall that in the third-party punishment game (§5.8) a first mover may transfer some of his endowment to a passive recipient, and then an observer (the third party) may devote some of his endowment to punishing the first mover, presumably in response to the third party’s disapproval of the first mover’s transfer. In the experiment all possible combinations of in- and out-group matches were implemented, so all three individuals might be from the same ethnic group, or just the first mover and the third party, or just the recipient and the first mover, or just the third party and the recipient. First movers gave more to in-group members, and this was true independently of the level of third-party punishment they anticipated. But significant amounts were given to out-group recipients, too. Stingy first movers were most heavily punished if the first mover was from one group and the recipient and the third party from the other. But third parties also avidly punished ungenerous first movers when all three were from the same group. And third parties also punished first movers from their own group even if the recipient was from the other, and even when both the first mover and the recipient were from a different group.

In subsequent one-shot trust game experiments (§5.11) with subjects also from two New Guinea highlands groups, the authors elicited players’ beliefs about the first mover’s expectations about the degree of reciprocity of the recipient (Efferson et al. 2010). When interacting with in-group members the amount transferred was uncorrelated with expectations of reciprocity, while transfers to out-group members correlated strongly with an expected
payback. Thus first movers appeared to be altruistic towards group members and strategic towards outsiders. But altruism towards outsiders was not entirely absent, nor was the third party indifferent to a first movers’s unfair treatment of an outsider. This is hardly surprising when one recalls that members of ancestral human groups not only fought one another, they also depended on one another for help in times of need, for information, mates, and trade goods. When we say that altruism is sometimes parochial, we mean that it recognizes group boundaries, not that it stops at the border. Modeling the evolution of these complex preferences in a way that may reveal how they could have evolved is the task we set for ourselves.

Inter-group aggression and in-group favoritism is similar to altruism in that it is often costly to the individual actor, who incurs mortal risks in a battle, or in shunning others, forgoes opportunities for beneficial coalitions, mating, co-insurance, and exchange. The struggle for survival of the Norse peoples who settled in Greenland for almost half a millennium prior to their demise around 1400 dramatically illustrates the perils of parochialism. Their hostile relations with the Inuit with whom they uneasily coexisted probably explains why they never learned the boat-building, fishing and hunting strategies that provided the basis for the Inuit survival there even through the Little Ice Age (McGhee 1984).

In the absence of positive assortment, neither parochialism nor altruism would seem likely to survive any selection process, whether cultural or genetic, that favors traits with higher payoffs. But parochial altruism could have emerged and proliferated among early modern humans because our ancestors lived in environments in which competition for resources favored groups with significant numbers of parochial altruists willing to engage in hostile conflict with outsiders on behalf of their fellow group members. These group benefits could have offset the within-group selection against both parochialism and altruism in a manner similar to that studied for simple altruism in the previous chapter. But the evolution of either altruism or parochialism singly seems unlikely, for as we saw in the previous chapter, intergroup hostilities favored the evolutionary success of individual altruism. A population of tolerant altruists would not fight their neighbors, while parochialism alone would have prompted frequent hostilities with few willing to fight. We thus have another co-evolutionary process but unlike the previous chapter where we studied the joint dynamics of culturally transmitted institutions and genetically transmitted behavioral predispositions, here we explore the coevolution of two individual traits, parochialism and
altruism, each providing the conditions for the evolutionary success of the other.

9.1 Parochial Altruism and War

With Jung Kyoo Choi we modeled a population of foragers who engage in both within- and between-group interactions in which individuals may adopt two types of behavior: altruism and parochialism (Choi and Bowles 2007). A related analytical population genetic model of “belligerence” (our parochialism) and “bravery” (our altruism), with results similar to the simulations reported here appears in Lehman and Feldman (2008). There are thus four behavioral types: parochial altruists, tolerant (non-parochial) altruists, parochial non-altruists, and tolerant non-altruists. Parochials (of either type) are hostile towards members of other groups. But only parochial altruists engage in combat, as the non-altruists do not bear personal costs in order to confer benefits on their fellow group members. In the absence of between-group hostilities, tolerant members of a group benefit from inter-group exchange, risk pooling and the kinds of mutually advantageous interactions with other groups mentioned in Chapter 7.

Two types of selection are at work in the model. Within-group selection favors tolerant non-altruists and tends to eliminate parochial altruists (as well as tolerant altruists and parochial non-altruists). By contrast, selective extinction resulting from inter-group conflict may favor parochial altruists. The reason is that if both hostility toward outsiders and a willingness to sacrifice on behalf of other members of one’s group are essential to being an effective fighter for the group, then groups with many parochial altruists will tend to win conflicts, and to replace groups with fewer.

In order to clarify the role of war, parochialism and selective extinction, we do not model the other mechanism by which altruism may spread, namely, selective emigration (Rogers 1990). Thus, in contrast to Maynard Smith’s haystack model and Wilson’s model of trait groups, in the absence of territorial expansion, highly altruistic groups in our model do not contribute more replicas to the next generation (Maynard Smith 1964, Wilson and Dugatkin 1997). Like the model in Chapter 8, our setting is quite unfavorable for the evolution of altruism as it is equivalent to models in which local density-dependent selection exactly offsets the group benefits of altruism (Taylor 1992, Wilson et al. 1992).
Parochial altruists do receive a direct benefit if a war occurs, as they share in their group’s increased probability of surviving a hostile encounter that results from their status as a ‘fighter’ (relative to the group’s likely survival had the individual been of another type.) For plausible group sizes and frequencies of conflict, however, this direct benefit is an order of magnitude smaller than the costs. So our parochial altruists are indeed altruistic: they would increase their fitness if they became tolerant non-altruists.

In every generation, at a cost of $c$, altruists ($A$’s) contribute to a public good whose value ($b$) is shared equally among the $n$ adult group members. The public good may be common defense and predation, or similar behaviors that confer benefits on all group members. Those who are not altruistic ($N$’s) do not contribute. Because we assume that $b > c > b/n$, contributing to the public good raises group-average payoffs but reduces the contributor’s payoffs and so is both group-beneficial and altruistic. Not contributing is the dominant strategy, the $N$’s payoffs exceeding the $A$’s by the amount $c$, irrespective of the distribution of $A$’s within the group. Payoffs in the absence of group conflict are described in Table 9.1.

<table>
<thead>
<tr>
<th></th>
<th>Parochials</th>
<th>Tolerant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altruist</td>
<td>$bf_i^A - c$</td>
<td>$bf_i^A - c + gn_j f_j^T$</td>
</tr>
<tr>
<td>Not Altruist</td>
<td>$bf_i^A$</td>
<td>$bf_i^A + gn_j f_j^T$</td>
</tr>
</tbody>
</table>

Table 9.1. Expected Payoffs to Four Behavioral Types in the Absence of Hostile Between-Group Interactions. The fraction of group $i$ who are altruists is $f_i^A$. All members receive the benefit of the public good, $bf_i^A$. Tolerant players of both types receive the benefits of non-hostile group interaction, $gn_j f_j^T$ where $g$ is the benefit of non-hostile group interaction, $n_j$ is the group size of the other group, and $f_j^T$ is the fraction of the other group who are tolerant. For our simulations, benchmark values of the parameters in the table are: $c = 0.01$, $b = 0.02$, $g = 0.001$.

In every generation each group interacts with another group either cooperatively, enjoying the exchange, coinsurance and other benefits of peaceful interactions described in Chapter 7, or in a hostile manner, as shown in Figure 9.1. Hostility in an inter-group interaction results if parochial members constitute a sufficiently large fraction of at least one group. The probability that a group survives a hostile interaction is increasing its number of fighters (parochial altruists) relative to the number of fighters in the opposing
group. The use of force between the two groups occurs when one is sufficiently likely to win, reflecting the fact that as with other primates, evenly matched human groups seek to avoid costly conflicts (Wilson et al. 2001). If a conflict occurs, a fraction of the members of the losing group are eliminated (the fraction eliminated is a constant times the between-group difference in the fraction of parochial altruists) and replaced by replicas drawn randomly from the winning group. Hostile inter-group interactions in our ancestral past may have more ongoing and less episodic than we model here, but taking account of this would not alter the causal mechanisms of the model.

![Figure 9.1. Between-group Interaction.](image)

If the interaction is not hostile, each tolerant ($T$) member receives a net benefit from each tolerant member of the paired group resulting from gains from trade or risk pooling. Parochials receive no benefits of this type. As a result, in the absence of war the expected payoff to the $T$’s in a given group exceed that to the $P$’s irrespective of the fraction of $T$’s in the group, so $T$ is the dominant strategy. This is why both $A$’s and $P$’s face adverse within-group selection.

The process by which intergenerational transmission of behaviors takes place could be cultural, those with higher payoffs being disproportionately
copied by the next generation, or genetic, payoffs measuring reproductive success. Every generation, members of each group are paired randomly to produce offspring, whose expected number is proportional to the parental couple’s share of the group’s payoffs.

So as not to favor the hypothesized coevolution of parochialism and altruism, which depends on the two behaviors being statistically associated, we assume an intergenerational transmission process with a strong tendency for the behaviors to be separated. Thus we assume no assortment in mating, so a parochial altruist is no more likely to mate with another parochial altruist than would occur by chance, and we also allow complete recombination, so that a parental couple composed of a PA and a TN will have offspring of all four behavioral types with equal probability. Additionally, this process is modified by mutation: with some probability $\mu$, each member’s offspring inherits a strategy randomly from the four possible types independently of the parental types. With probability $1 - \mu$, the non-mutational replication described just above takes place. Each generation, with some probability $(m)$ each member migrates to a randomly selected group.

This replication process reflects the assumption that all sites are saturated so that a group’s population can grow only if it increases its carrying capacity by territorial aggrandizement.

### 9.2 Results

The agent-based computer simulation explored its properties of this model under a range of parameters calibrated to resemble the environment of late Pleistocene and early Holocene humans. Figure 9.2 shows that over a very long period, the simulated population spends most of the time in states with many parochial altruists and few of the other three types (the upper-right cluster of bars), or in states with many tolerant non-altruists and few of the other three types (the lower-left cluster). In the former case, high levels of parochialism in the population sustain a high level of conflict among groups, thereby making between-group selection a formidable evolutionary process and as a result maintaining a substantial fraction of parochial altruists in groups. This replicates the synergy between war and altruism in Chapter 8 (see Figure 8.6). By contrast, when the population is in the states in which tolerant non-altruists are prevalent, few wars occur. As a result the within-group selection pressures against parochials and altruists predominate, maintaining only a low level of each.
Figure 9.2. Parochial-altruist and Tolerant Non-altruist Outcomes Occur with High Frequency. The parameter values are as described in the legend of Table 9.1. The height of the bars gives the fraction of a very long period in which we observe the indicated pair of population level frequencies of altruists and parochials in the population. These frequencies are an approximation of the stationary distribution of the underlying stochastic Markov process implied by our model and have been recovered from a very large number of implementations of the model with initial seeds at every state in the state space. The baseline parameters from Table 9.1 are used in this simulation, along with migration and mutation rates $m = 0.3$, and $\mu = 0.005$. The average group size is 26 members per generation.

In this model, parochialism and altruism evolve not because the parochial altruists deliberately associate with like types, but rather because within-group interactions for which cooperation is possible are characterized by positive assortment. Because parochial altruists are more likely to be in groups with other parochial altruists than under population-wide random matching, while hostile interactions are characterized by negative assort-
ment because the wars in which most parochial altruists participate (and win) tend to be against groups with larger fractions of the other three types. This is a result of the fact that evenly matched groups tend not to fight. When the number of parochials in the population is sufficient that such hostile conflicts are frequent, the result is to favor the proliferation of both altruists and parochials. Similarly, when tolerant individuals are prevalent in the population they benefit from positive assortment in cooperative interactions because most are in groups that reap benefits from peaceful intergroup relations.

The arrows in Figure 9.3 give the direction of expected movement for a population whose composition is indicated by the root of the arrow. The longer arrows indicate strong selection against parochials in the absence of significant frequencies of altruists and against altruists in the absence of significant frequencies of parochials.

The extent and nature of inter-group relationships are produced endogenously by the model. Thus a comparison of fatalities in war and other aspects of inter-group interactions with available empirical information provides a check on the plausibility of the model and parameters chosen. Tolerant individuals (T’s) on average receive benefits from non-hostile interactions with all tolerant members of other groups amounting to 0.016 for all states. These values may be compared to the benefit of 0.02 that each altruist confers on the members of his own group by contributing to the public good. There are few benefits to T’s near b in Figure 9.3 because at this state most interactions are hostile and even in the peaceful interactions there are few T’s in the other group. Near b, a group may expect to be engaged in a war every 7.1 generations, the probability of war in each generation being 0.14, if it loses the war, to suffer fatalities equal to about 40 percent of its population, and if it wins, to suffer fatalities of about 10 percent, who are all PAs. In the neighborhood of b, about 3.6 percent of the entire population dies in warfare per generation. Thus both altruism and parochialism are sustained by levels of inter-group conflict and deaths in warfare considerably below the estimates from archeological and ethnographic data presented in Chapter 7. We are therefore quite confident that our results do not require implausibly high levels of between-group hostility. Figure 9.4 illustrates the transition process between states close to a and those at b.
Figure 9.3. **Probabilistic Transition Dynamics.** The parameter values are as before. Each arrow represents the expected change at each state, based on a transition matrix recovered from the underlying Markov process as described in the previous figure legend. Asymptotically stable states occur where both frequencies are approximately 15 percent (point a) and in the neighborhood of the frequencies of A and P equal to (0.95, 0.75), and point b. Point c is a saddle (unstable critical point).
Figure 9.4. **Transitions, Peace and War.** The top panels illustrate transitions between states similar to a and b in the previous figure. The middle and bottom panels show that high levels of P and A sustain high frequencies of warfare and vice versa. Wars are most frequent when about half the population are parochial altruists, because at these frequencies imbalances between groups are more common.
In Choi and Bowles (2007) we explore the sensitivity of these results to variations in the parameters. The frequency of parochial altruists and of war varies inversely with group size and the migration rate because increasing these population structure parameters diminishes the between-group differences in the distribution of types, thereby weakening the effects of selective extinction. The parameters affecting within-group and between-group selection also have the expected effect: war and parochial altruism vary positively with the extent of losses inflicted on losers and inversely with both the benefit from non-hostile interactions ($g$) and the cost of altruism ($c$).

The explanation of the evolution of parochial altruism outlined here also provides a plausible explanation of the results of the experimental third-party punishment experiment in Papua New Guinea mentioned at the beginning of the chapter. Suppose that the experimental subjects pool comprised the four behavioral types in our model. Tolerant altruists are expected to bear costs in order to give to both insiders and outsiders, and to punish those who violate norms. In view of the importance of mutually beneficial intergroup relations, punishment of norm violators by altruists would include out-group members as well as insiders. But parochial altruists would give preferentially to ingroup members and punish those who harm group members more severely than if the victim is not an insider. Our model shows that this spiteful behavior towards outsiders could have evolved by benefitting other group members in hostile intergroup contests. In the experiment the punishment of an outsider increases the relative payoffs of the actor’s group as the cost to the target of punishment is assumed to be three times the cost to the punisher.

Other evolutionary influences may also be suggested by the experiment: the fact that group members are likely to be more genetically related to insiders than to outsiders might be part of the explanation of group favoritism in giving. But this could not explain giving itself in a non-repeated interaction. The reason is that in the experiment the benefit conferred by the gift is fully offset by the cost to the giver and so could not have evolved by kin selection (that is, $b = c$ in the terms of equation 3.1). Hamilton’s rule in this case would require that for giving not to be inclusive fitness reducing the giver and receiver would have to be genetically identical, and even then the giving would be simply selectively neutral! Moreover, neither direct fitness benefits nor inclusive fitness can explain the punishment of norm violators, or the more severe punishment of outsiders, assuming the ances-
tors of these experimental subjects lived in groups of the general size of our foraging forebears.

9.3 Persistent Parochialism

Though its origins thus may lie in Pleistocene and Holocene warfare, the persistence of parochial hostility to “outsiders” in modern politically liberal and economically competitive societies is something of a puzzle, because in addition to the disabilities associated with limited gains from trade and economies of scale, these sentiments are often condemned for the racism, religious intolerance and other repugnant forms that parochialism sometimes takes.

To address this puzzle we studied modern parochial networks, defined as sets of individuals unified by similarity of one or more ethnic or other ascriptive characteristics and engaged in non-anonymous interactions structured by high entry and exit costs, but lacking a centralized authority. This work, (Bowles and Gintis 2004), draws on the works of Homans (1958), Ben-Porath (1980), Iannaccone (1992), Kandori (1992), Greif (1994), Akerlof (1997), Pagano (1995), Bénabou (1996), Durlauf (1996), Kranton (1996), Taylor (1997) and Glaeser (2000). In this model, parochial sentiments persist, despite the associated costs, because within-group cultural diversity may make the informal enforcement of contracts and other aspects of cooperation more difficult. It thus bears a strong affinity with the class of economic models of Phelps (1972), Arrow (1973), Sethi et al. (2010), Bénabou (1993) and others who sought to understand how group differences could persist in a politically liberal and economically competitive environment.

Our model may help explain why modern-day parochial networks often do quite well economically, as the flourishing informal ethnic business linkages among new immigrants to the United States and the United Kingdom attest (Rauch 1996, Granovetter 1985, Kotkin 1993). For instance, Cambodians run more than 80 per cent of California’s doughnut shops, raising funds from friends, family, and ethnic credit associations (Kaufman 1995). Similarly, Indians own more than a third of the motels in the United States, frequently raising initial capital through unsecured loans from extended family members (Woodyard 1995).

When interactions among group members are characterized by material payoffs that take the form of prisoner’s dilemma, public goods game or
other social dilemmas, parochialism may contribute to successful cooperation in three ways. First, members may feel more altruistic toward the ethnically similar members of their group than towards “outsiders.” As we have seen in Chapter 5, if each member values the payoffs gained by other members sufficiently highly, mutual defect is no longer the dominant strategy equilibrium of these interactions, and mutual cooperation may be a stable Nash equilibrium.

A second way that cultural affinity could support cooperation is by enhancing the force of altruistic punishment that often sustain cooperation in experimental public goods games. For example, it is plausible that the shame induced by criticism for selfish behavior, studied in Chapter 12, is greater when the defector shares the critic’s norms about good behavior.

The third reason parochial groups may cooperate in interactions that would take the form of a prisoner’s dilemma or public goods game in non-parochial settings, unlike the first two reasons, does not concern the members’ preferences. Rather it is the effect of cultural affinity on the information structure of the interaction, allowing equilibrium strategies unavailable in the information environments of less parochial groups. This third reason for the success of parochial networks receives strong support the remarkable experiments conducted in the slums of Kampala described in Chapter 5. In these communities there is considerable evidence that more ethnically homogeneous neighborhoods sustain higher levels of cooperation than do more heterogeneous neighborhoods. The authors (Habyarimana et al. 2009) report that:

> successful collective action among homogeneous ethic communities...is attributable to the existence of norms and institutions that facilitate the sanctioning of non-contributors. We find no evidence for a commonality of tastes within ethnic groups, for greater degrees of altruism towards co-ethnics or for an impact of shared ethnicity on the productivity of teams.

Our analysis of parochial networks explores this superior sanctioning and norm enforcement capacities of parochial groups and their resulting success in surmounting social dilemmas. But like parochial foraging groups, modern parochial networks generally face offsetting costs including foregone gains from trade and economies of scale and the inability to exploit productive diversity (Page 2007), as when different groups have complementary skills or face uncorrelated shocks, making partnerships for cross-
ethnic production or co-insurance mutually beneficial. The advantages of trade with those deemed “outsiders” is a common explanation of the permeability of network boundaries in small scale societies (Adams 1974) and of the extinction of very restrictive networks in favor of more inclusive entities (Gellner 1985, Weber 1976). Ben-Porath (1980) develops similar reasoning concerning the economic capabilities of families and other face to face networks:

The transactional advantages of the family cannot compensate for the fact that within its confines the returns from impersonal exchange and the division of labor are not fully realizable. (p. 14).

A particularly well-documented example of this tension is Greif’s (1994) account of how the competitive advantages stemming from the superior within-network contractual enforcement capabilities of the tight-knit 13th century community of Maghribi merchants in the Mediterranean was eventually offset by their reduced ability to engage in successful exchange with outsiders, resulting in their inability to compete with the more individualistic Genovese traders.

People, of course, do not normally express their identification with parochial networks in terms of their economic advantages. Rather, they typically invoke religious faith, ethnic purity, or personal loyalty. The contribution of parochial sentiments and practices to economic performance of groups is far from the sole reason for their persistence. Ethnic, racial and other group identities arise and persist for a multitude of reasons, many of them far less benign than those studied here. Loury (2001) and Glaeser (2005) provide accounts of some of these reasons. Socialization into parochial preferences is commonly part of the oblique cultural transmission process studied in Chapter 11.

9.4 The Legacy of ‘Red in Tooth and Claw’

Our approach thus explains *Homo sapiens* as a war-like species and the status of parochial altruism as a common and enduring suite of human behaviors. It also shows that warfare itself may have contributed to the spread of human altruism. We initially recoiled at this unpleasant and surprising conclusion. But the simulations and the data on prehistoric warfare tell a
convincing story, one that is corroborated in a non-evolutionary setting by a tragic natural experiment.

Maarten Voors and his coauthors (2010) studied the massive indiscriminate killing that took place during the genocidal violence in Burundi and the effects that it had on individuals’ altruistic behavior towards neighbors. A total of 300 subjects from 35 randomly selected villages, 26 of which had experienced lethal ethnic violence over the period of Hutu-Tutsi conflicts during the years 1993-2003, participated in an experiment in which they could allocate money between themselves and another anonymous individual from their village. Controlling for a number of demographic and community differences, those from villages that had been exposed to ethnic violence were much more likely to behave altruistically towards their neighbors in the experiment. The same was true of the individual experience of violence. Those who had themselves been victim of an attack were also more altruistic. The authors provide persuasive reasons to believe that the direction of causation is from intergroup conflict to within group altruism, and not the other way around (Voors et al. 2010). These surprising effects of intergroup violence on experimental behavior were substantial. A standard deviation difference in the number of attacks suffered by a village was associated with almost a half a standard deviation increase in altruistic behavior. Exposure to violence was also significantly related to greater participation in community organizations.

But was warfare necessary in the evolution of human altruism? Could altruism not have evolved in the absence of lethal group conflict, and perhaps taken a tolerant rather than parochial form? The key element in the explanation offered here and in Chapter 6 is that groups with more altruists survive challenges, encroaching on less cooperative groups or even eliminating them. The differential survival of more cooperative groups need not have been the result of warfare. The tumultuous climate of the Late Pleistocene (Figure 7.1) presented groups with extraordinary challenges quite apart from direct confrontations with other groups. If groups with a greater number of altruists were more likely than less cooperative groups to surmount these environmental challenges they would have had opportunities to occupy the vacated sites of the defunct groups of less altruistic and less cooperative individuals, replicating the process modeled here as the result of warfare. Thus, returning to Darwin in the head quote of the previous chapter, “the social and moral qualities” could have “advanced and dif-
fused throughout the world” entirely without any tribe being “victorious over other tribes.”

While altruism might have evolved in the absence of group conflict, we do not think that it did. The evidence for frequent lethal encounters and the plausibility of the assumption that more cooperative groups would survive these challenges suggests that warfare, no doubt in conjunction with environmental challenges, played a critical role in the evolution of a cooperative species.

But a form of altruism circumscribed by parochialism is in our legacy it need not be our destiny. The fact that altruism and parochialism may have a common evolutionary origin, whether cultural or genetic, does not mean that the two are inseparable. Examples of tolerant, even anti-parochial, altruism include some inter-group behavioral experiments (Bouckaert and Dhaene 2003), the electoral support in many countries for tax-supported economic aid to the people of poor nations, and the participation of people of all ancestral groups in political movements against racism.

The deliberate processes of socialization and learning we study in Chapter 11 are capable of supporting tolerant altruistic behavior.
10

The Evolution of Strong Reciprocity

My motive for doing what I am going to do is simply personal revenge. I do not expect to accomplish anything by it.... Of course, I would like to get revenge on the whole scientific and bureaucratic establishment... but that being impossible, I have to content myself with just a little revenge.


The previous two chapters showed that an unconditional form of altruistic cooperation among members of a group could have evolved under conditions likely to have been experienced by our late Pleistocene and early Holocene ancestors. But we also know that in experiments and everyday life, altruism is rarely unconditional. We have seen, for example, that when no other recourse is available, altruistic cooperators react to free-riding by others by withdrawing their contributions. But a cognitively advanced animal can do a lot better than to just walk away. He and his fellow group members can gang up on the defector.

Together, a predisposition to cooperate and a willingness to punish defectors is what we have termed strong reciprocity, and it is the combination of the two that accounts for the large-scale cooperation exhibited by our species. Here we show how a willingness to punish those who violate social norms even at a cost to oneself could have evolved.

Punishing defections has an advantage over simply withdrawing cooperation. As we saw in Chapter 3, withdrawing cooperation indiscriminately harms all group members, whereas punishing defections is targeted at the miscreants. Moreover, when the willingness to punish is common, all defections are due to mistakes or exceptional conditions, such as sickness. Therefore when a group is composed of a sufficient number of punishers, the disadvantage of the Punishers compared to Nonpunishers becomes quite small, since they rarely incur the cost of punishing. Thus once strong reciprocators are common in a group, the higher average payoffs of the more cooperative groups in which most strong reciprocators find themselves can readily offset their occasional direct costs of punishing the occasional de-
In the Price equation (6.2) this means that the cost to the individual strong reciprocator, $c_i$, is small, so that neither the group-level benefits of cooperation nor the between-group differences that make up the between-group selection term need to be very large in order to stabilize the fraction of strong reciprocators in a population.

While punishers may be greatly disadvantaged when rare, making their initial emergence difficult, once established, punishers can be maintained in a population as a result of simple group competition. Moreover, if punishers are clever enough to refrain from incurring the costs of punishing when punishing is not cost-effective, such as when there are many Nonpunishers and few punishers, their disadvantage when rare is also very small.

There are two important problems with this explanation. First, punishment is costly to both the punisher and the target and can be counterproductive if the costs of punishing exceed the resulting gains from cooperation (Fehr and Gächter 2000a, Bochet et al. 2006, Cinyabuguma et al. 2006, Herrmann et al. 2008). This problem is exacerbated when punishers target cooperative group members, as sometimes occurs in experiments. Second, the initial emergence of punishment remains a puzzle. In order to survive, punishers must engage in enough punishment of defectors so that the induced cooperation more than offsets the cost of punishing. When punishers are rare they must bear this cost alone rather than sharing it with other punishers. (Boyd and Richerson 1988, Panchanathan and Boyd 2003, Boyd and Richerson 1992, Boyd et al. 2003).

These problems are an artifact of the unrealistic way that punishment is implemented in existing models and in most experiments. In these models, punishment is an unconditional and uncoordinated individual action automatically triggered by defection. Similarly, with few exceptions in experiments (Ostrom et al. 1992), individuals cannot coordinate their punishment. As we saw in Chapter 2, the ethnographic evidence indicates that punishment is coordinated by communication among punishers, is contingent on the expected effectiveness of punishment in inducing cooperation, and is not undertaken unless it is judged as legitimate by most group members (Wiessner 2005, Boehm 1993, Mahdi 1986). As a result, solitary individuals rarely attempt to punish those who violate social norms. When it occurs, punishment is usually collective and conveys a message of peer condemnation. Consistent with the anthropological evidence, in behavioral experiments with communication or with a choice of a collective punish-
Punishment strategy by subjects, punishment is often effective in raising group average payoffs (Ertan et al. 2009).

Here, we use both an analytical model adapted from Boyd, Gintis, and Bowles (2010) and an agent-based model with two empirically-grounded features that resolve the above difficulties. First, punishment is coordinated among group members so that it is contingent on the number of others predisposed to participate in the punishment. This is an example of what biologists call “quorum sensing,” practiced by bacteria and other organisms (Miller and Bassler 2001, Diggle et al. 2007). When punishers are also quorum sensors, then when they are rare, they demur and so bear only the minor cost of signaling their willingness to punish. They thus avoid the cost of punishing when it does not pay. Second, consistent with the “strength in numbers” and “divide and rule” maxims, punishment is characterized by increasing returns to scale, so the total cost of punishing declines as the number of punishers increases. Adding these two novel features resolves the problems with previous models. Our model shows that for levels of relatedness consistent with recent genetic data from hunter-gather populations, punishment can proliferate when rare, and when it is common, it increases group-average fitness.

10.1 Punishment with Quorum Sensing

Consider a large population in which individuals interact repeatedly in groups of size $n$. Groups are randomly formed so there is no genetic assortment. Later, we will introduce an empirically plausible degree of genetic assortment among groups. The initial period in the life of a group has three stages. First is a **signaling stage** (at cost $q$) in which individuals can signal their intent to punish any defector. The cost of signaling is high enough that it does not pay to signal and then fail to punish. There follows a **cooperation stage**, during which individuals can choose to cooperate or defect. Cooperation costs the cooperator $c$ and benefits each member of the group $b/n$ ($b > c > b/n$).

Finally, there is a **punishment stage** in which individuals can coordinate with other punishers to administer punishment. To model the punishment process, we use a probabilistic version of Lanchester’s Law on the relationship between numbers on opposing sides of a conflict and success in contests as the basis for our expected cost of punishment for those engaging in a punishment episode against a target (Lanchester 1916, Engel 1954, Hwang
The expected cost of engaging in a punishing party depends on the likely outcome of the encounter with the target.

We take account of this aspect of punishment in the following way. Suppose the cost of punishing, \( k \), is borne by a randomly selected member of the group of Punishers if and only if the punishing episode is a standoff, a situation in which neither the Punishers nor the target “win” so that both target and Punishers bear costs.

Suppose further that a standoff occurs with probability \( 1/n_p \), where \( n_p \) is the number of Punishers against a lone target. Thus, for instance, a single Punisher against a single target always results in a standoff. The expected cost of joining a group of \( n_p \) Punishers is thus the probability of a standoff \( (1/n_p) \) times the expected cost that a member of the punishing party will bear in the case of a standoff \( (k/n_p) \), giving expected cost \( k/n_p^2 \).

During subsequent periods, there is not signaling stage, as players already know the number of Punishers in the group. The cooperation and punishment stages, however, remain, after which the interaction continues to another period with probability \( \delta \), so the expected duration of a group is \( 1/(1 - \delta) \) periods. When all groups have disbanded, new groups are drawn from the population and the process is repeated.

Population structures like this one, in which groups do not persist forever, but rather are created anew periodically by drawing individuals from a larger population, are widely used in models of social evolution, including models of repeated interaction that lead to the evolution of contingent behavior (Axelrod and Hamilton 1981, Nowak and Sigmund 1998b). Such models provide an analytically tractable approximation to more realistic structures in which extensive gene flow among persistent groups causes low relatedness within groups. In the first period of such models individuals have no common history, as they would if we modeled persistent groups, and hence cannot know anything about strategies of other group members. To address this information problem it is standard to introduce a first information gathering period in which individuals may behave differently than in subsequent periods. This is a seemingly unrealistic assumption. But even in the more realistic setting of persistent groups, individuals change, die or leave the group and are replaced by migrants or offspring. This means that actors must in any case deal with situations in which the past behavior of some group members is unknown, analogous to the first period in the present model. We believe that the present model represents a worst case for the evolution of punishment because it maximizes the level of uncertainty.
about the strategies of others and hence maximizes the cost to punishers of determining if a sufficient number are present to warrant punishing.

Individuals have one of two heritable strategies, Punisher and Nonpunisher. Later we test the robustness of our model by introducing Liars, who signal their willingness to punish but then do not, and Opportunists who signal, but punish only if their participation is needed to secure a quorum.

During the first period in the life of a group, Punishers signal they are willing to punish, whereas Nonpunishers do not. Next, if at least \( \tau (0 < \tau < n - 1) \) other group members signal, Punishers cooperate with probability \( 1 - \epsilon \) and defect with probability \( \epsilon \), and then punish any individual who did not cooperate. We refer to Punishers with a threshold of \( \tau \) as \( \tau \)-Punishers. If fewer than \( \tau \) other individuals signaled during the first stage, Punishers defect, and do not punish. Nonpunishers do not signal, defect, and do not punish, and as a result are punished if there are at least \( \tau + 1 \) Punishers in the group. During subsequent stages, both types cooperate with probability \( 1 - \epsilon \) and defect with probability \( \epsilon \) if defectors were punished the last time a defection occurred. Punishers punish defectors if at least \( \tau \) other individuals punished the last time a defection occurred.

The cost of being punished to the target, \( p \), is greater than the net cost of cooperating, \( c - b / n \), so, on average, cooperation is the payoff maximizing action if punishment is anticipated. A fraction \( e \) of individuals nonetheless defects, either due to error, or because cooperation is more costly for some individuals and so it does pay them to cooperate, even if they expect to be punished. Nonpunishers are a plausible ancestral state for the evolution of punishment. They do not cooperate or punish, nor do they respond to unverified threats. However, once they have been punished, they cooperate in subsequent periods in order to avoid more punishment. Note that cooperation is a facultative choice, not an inherited behavior. It has been argued that punishment can evolve only when it is linked to cooperation (Lehmann et al. 2007). After the first period, Punishers and Nonpunishers cooperate under exactly the same conditions, so the linkage between cooperation and punishment is extremely weak. In Boyd, Gintis, and Bowles (2010), we show that even this weak linkage is not necessary for the evolution of punishment. After the social interaction just described, individuals reproduce at a rate that is proportional to their payoff compared to the population-average payoff leading to the equations that describe how natural selection changes the frequencies of the two types through time.
Figure 10.1. Equilibrium frequencies of Punishers for Two Values of $b$. The quorum is $\tau + 1$ when group members are unrelated. For each value of $\tau$ the solid circles give locally stable equilibrium frequencies of the punishing type, and the open circles give interior unstable equilibrium frequencies. For Panel A, with $\tau < 3$, the only stable equilibrium is a population without Punishers. For larger $\tau$, there are two stable equilibrium frequencies, zero, and a mixed strategy at which Punishers and Nonpunishers coexist. The unstable equilibrium marks the frequency that Punishers must achieve before they are favored by selection. For Panel B, there are two equilibria for all values of $\tau > 0$. Benchmark parameters: $c = 0.01$, $q = k = p = 1.5c$, $r = 0$, $a = 2q = k = p = 1.5c$, $\epsilon = 0.1$, $n = 18$, $T = 25$.

To explore the workings of the model, these values are varied one parameter at a time from the baseline parameter set listed in Figure 10.1. Before presenting results based on groups formed for plausible levels of between-group genetic differences, we consider the case of randomly formed groups (namely, the degree of relatedness of group members is $r = 0$).

For most parameter sets, there are two long-run evolutionary outcomes, illustrated in Figure 10.1, which shows the equilibrium frequency of Punishers as a function of the level of the quorum threshold. For example the left panel shows that if the threshold is 8, then populations with no Punishers and with 60% Punishers are both stable equilibria (the solid dots) and that in populations with fewer than 24% punishers (the open dot) they will
be eliminated, while in populations with more than 24% Punishers, their numbers will expand to 60% of the population. Technically the open dot gives the boundary of the basin of attraction of the two stable equilibria, much like point c in Figure 9.3. Figure 10.1 shows that for any threshold level, universal non-punishment is stable. When Punishers are rare, they will be alone in a group and thus pay the cost of signaling but do not reap the benefits of cooperation, and thus will have lower fitness than Nonpunishers. If Punishers are willing to punish alone (τ = 0), then they can invade a population of all Nonpunishers but only if a single Punisher can recoup the costs of signaling and punishing. Given our parameters, this rather implausible condition is not satisfied, so for all of our models, only cooperative punishment pays (τ ≥ 1).

Second, given our parameters, for all but the lowest thresholds, mixtures of cooperative Punishers and Nonpunishers can also be evolutionary stable, as indicated by the solid circles in the interior of Figure 10.1. The existence of a stable mixture of Punishers and Nonpunishers depends on the value of the punishment threshold, τ. When the threshold is too low, punishment does not pay in the threshold group, and non-punishment is the only stable equilibrium. At higher thresholds, punishment does pay in threshold groups, and this means that punishment can be favored if such groups are sufficiently common. Thus, as the frequency of Punishers in the population increases from zero, the fraction of groups with the threshold number of cooperators increases, and so does the fitness of Punishers, as seen in Figure 10.2. Once the fraction of groups with the threshold number of cooperators is sufficiently high, Punishers’ advantage in threshold groups offsets their disadvantage in all other groups (the rising function in Figure 10.2 crosses the zero line). For frequencies of Punishers greater than this, natural selection will increase the frequency of Punishers. The zero-line intersection marks the unstable equilibrium shown by the open dots in Figure 10.1. Further increases the frequency of Punishers, eventually decrease the fraction of threshold groups. When the fitness of Punishers and Nonpunishers is again equalized (where the falling curve intersects the zero line), there is a stable, mixed equilibrium. As τ increases, the frequency of Punishers at the mixed equilibrium also increases, but the basin of attraction of this equilibrium decreases. Note that at the equilibrium, punishment is not altruistic. Averaged over all groups, the long run benefits of punishment exactly compensate for the costs. However, it is mutually beneficial to the
Figure 10.2. **Fitness Differences.** The difference in fitness of (unrelated) Punishers ($W_P$) and Nonpunishers ($W_N$) as a function of the frequency of Punishers. When this difference is positive, Punishers increase in frequency, and when it is negative Punishers decrease in frequency. Equilibria occur when this difference is zero. The equilibrium is stable when the function intersects the horizontal axis from above and unstable otherwise. When $\tau = 1$, punishment at the threshold does not pay for any frequency of Punishers, and thus increasing the frequency of Punishers from zero decreases their relative fitness. For larger values of $\tau$, punishment at the threshold does pay, and thus increasing the frequency of Punishers increases their fitness. This leads to a stable mixed equilibrium at which Punishers and Nonpunishers coexist. Parameters as in Figure 10.1.

While punishment is stable in this model, so is non-punishment. A complete account of the evolution of cooperation must explain how punishing strategies can increase when rare. In their classic work on pair-wise reciprocity, Axelrod and Hamilton (1981) suggest that a small amount of non-random assortment, such as interaction between weakly related group members, destabilizes uncooperative equilibria, but not cooperative ones.
Subsequent work has shown that this principle holds in a wide range of pairwise cooperative interactions, but not in interactions in larger groups (e.g. Boyd Richerson 1988, 1992, Panchanathan and Boyd 2004, Gardner and West 2004). The major exception to the this conclusion is Hauert et al. (2007) who show that the addition of non-participants who neither cooperate nor consume the benefits of cooperation allows the invasion of Punishers, but only under a limited range of conditions (Boyd and Matthew 2007).

In our model, modest amounts of assortative interaction allow cooperative Punishers to increase when rare even in quite large groups. To allow for assortment, the relatedness within groups varies. When \( r = 0 \), groups are formed at random, as in our results thus far. When \( r > 0 \), individuals are more likely to be in groups with individuals like themselves. Figure 10.4 shows the equilibrium behavior of the model assuming that \( r = 0.07 \) a degree of genetic differential a lot less than the average relatedness with in human foraging groups (Table 7.1). Now, for low thresholds (\( \tau \leq 4 \)) the only stable equilibrium is a mixture of Punishers and Nonpunishers, so

---

**Figure 10.3. Average Fitness Differences** The difference in average fitness between the mixed equilibrium at which Punishers are present and the monomorphic nonpunishing equilibrium. Whenever the mixed equilibrium exists it has higher average fitness, but near maximum benefit differences occur for relatively low thresholds. Parameters as in Figure 10.1.
Punishers can invade when rare. As is shown in Boyd, Gintis, and Bowles (2010), this result persists when groups are much larger ($n = 50$) and for lower levels of relatedness if the benefit cost ratio is somewhat higher. Interestingly, we also show in Boyd, Gintis and Bowles (2010) that modest assortment does not allow Punishment strategies with higher thresholds to invade populations with Punishers with lower thresholds. Thus the model predicts that only some individuals will engage in cooperative punishment.

The results presented so far depend critically on the cost of signaling the willingness to punish, $q$. Punishers honestly signal their willingness to punish and thus avoid the cost of punishment when there are too few Punishers. To determine the minimum cost of signaling necessary to insure
that the signal is honest we introduce a third strategy, Liar. These are the second-order free-riders. The Liar signals in the signaling stage but never punishes. Like the Punisher, the Liar cooperate in any period if and only if the quorum is met. The cost of the Liar strategy is the signaling cost \( q \). The gain is the number of periods in which the quorum appears to be met but in fact is not. Because of errors, this is likely to be no more than a single period. Thus if \( q \) is sufficiently large, the Liar cannot invade a population of Punishers and Nonpunishers, or Punishers and Contingent Cooperators (players who mimic Nonpunishers except that they cooperate on the first round when the quorum is met). We show in Boyd, Gintis and Bowles (2010) that \( q = p \) is sufficient to prevent invasion by Liars for plausible parameter values.

These results are sensitive to the assumption that there are increasing returns to scale in punishment. When \( a = 1 \) the total cost of punishing defectors is independent of the number of Punishers. The results are quite different. Much higher frequencies of punishment are required before punishment becomes stable. This supports the intuition that increasing returns is crucial, and therefore, the notion of coordinated punishment is important.

### 10.2 Altruistic Punishment in a Realistic Demography

In this section we present an agent-based model extending the analysis of the previous section, adding some demographic realism. We here assume that the larger population persists for a great many generations (typically 100,000 or more), and when a group is formed, it lasts throughout the history of the population unless its size becomes too small to sustain life under Pleistocene conditions. We set the average and minimum group sizes at 30 and 8 members, respectively. Individual group size and composition change through birth, death, and migration, although we maintain a constant number of groups and a constant total population. The type of a new group member becomes public information when the individual does or does not signal his type as Punisher, at cost \( q \). We assume behavior is the expression of genetic inheritance and individuals are haploid (an individual has one copy of each gene), but reproduction is diploid (an individual inherits each gene with equal probability from one of two parents). Finally, we assume \( \tau = 6 \) although, a smaller value would have led Punishers to invade a Nonpunisher population more rapidly.
We will later add a new type, the Opportunist, who signals but only pun-
ishes if he is needed to attain a quorum. The Opportunist clearly domi-
nates both the Liar treated in the previous section, and the Cooperator, who
does not signal or punish, but who cooperates unconditionally in all peri-
ods. We will see that the Opportunist is driven to extinction by Punishers,
and a for t i o ri, the same fate would be in store for Liars and unconditional
Cooperators.

We assume that in the first period following the group’s formation, all
Punishers bear a cost of $q$ per member to determine the member’s type, and
this is then common knowledge for all members of the group. In succeed-
ing periods, migration, birth, and death will lead to the disappearance of
some Punishers and the appearance of others individuals (immigrants and
new-borns) who we assume are of unknown type. We assume that when the
number of known Punishers is $\tau$ or fewer, but the number of known Pun-
ishers plus individuals of unknown type is greater than $\tau$, the individuals of
unknown type who are Punishers signal their willingness to punish at cost
$q$.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>2</td>
<td>Increasing Returns to Scale in Punishing</td>
</tr>
<tr>
<td>$b$</td>
<td>0.04</td>
<td>Benefits of Cooperation</td>
</tr>
<tr>
<td>$c$</td>
<td>0.01</td>
<td>Cost of Cooperation</td>
</tr>
<tr>
<td>$k$</td>
<td>0.015</td>
<td>Cost of Punishing</td>
</tr>
<tr>
<td>$p$</td>
<td>0.015</td>
<td>Cost of Being Punished</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.04</td>
<td>Reproduction Rate per Period</td>
</tr>
<tr>
<td>$\mu$</td>
<td>0.001</td>
<td>Mutation Rate</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>0.03</td>
<td>Execution Error Rate</td>
</tr>
<tr>
<td>$m$</td>
<td>0.003</td>
<td>Migration Rate per Period</td>
</tr>
<tr>
<td>$m'$</td>
<td>0.30</td>
<td>Fraction of Newborns who Migrate to Other Groups</td>
</tr>
<tr>
<td>$\tau$</td>
<td>6</td>
<td>Threshold Number of Punishers</td>
</tr>
<tr>
<td>$n$</td>
<td>30</td>
<td>Average Group Size</td>
</tr>
<tr>
<td>$n'$</td>
<td>8</td>
<td>Minimum Sustainable Group Size</td>
</tr>
<tr>
<td>$W_0$</td>
<td>0.95</td>
<td>Baseline Fitness</td>
</tr>
</tbody>
</table>

Table 10.1. **Baseline Parameters.** These parameters are used in all simulations, unless otherwise noted.
We assume 25 production periods occur per generation, so the periods are roughly years. This corresponds to \( T = 25 \) and \( \delta = 0.96 \) in the analytical model. Each individual mates and reproduces on average twice per generation, so the per-period reproduction rate is \( \rho = 0.04 \). The baseline parameters are listed in Table 10.1.

There are 250 groups, giving a total population of 7500. Because of birth, death, and migration, group size is variable over time, but total population is fixed at 7500. We assume that 70% of newborns are located in the parental group, and 30% are relocated randomly in the population. This implies a total migration rate of \( 0.3 + 0.003 \times 25 = 37.5\% \) per generation, or 1.5% per year. This allows for exogamy plus some random migration. Moreover, we assume that when a group becomes sufficiently small, it is repopulated by additional migration from larger groups and more successful. This repopulation is a form of migration that supplements other forms of migration. Consistent with Pleistocene demographics (Hassan 1973), we study a population with constant size, so the death rate is 4% per period, and new individuals are born at the rate of 4% per period as well. To avoid the artificial situation in which Punishers bear no cost because all individuals cooperate, we assume that with 3% probability, an individual who attempts to cooperate fails, and in fact defects. This means that in a group of thirty at least one defection will occur in a given period with 60% probability.

We start each simulation with only Nonpunishers in the population, so Punishers only enter a group through the random mutation of newborns. We assume offspring mutate at rate 0.001 from the parentally supplied genotype. Whatever the parental type, the mutant is Nonpunisher with probability 0.97, and is a Punisher with probability 0.03. Thus, in every 100 generations, there are on average 750 mutants, of which \( 750 \times 0.03 = 22.5 \) are Punishers. We normalize payoffs by a linear transformation so that the payoffs \( \pi \) vary from zero to one. We assume a baseline fitness \( W_0 = 0.95 \) and a selection coefficient \( 1 - W_0 = 0.05 \) on \( \pi \), so that individual fitness is given by \( W_0 + (1 - W_0)\pi \), normalized so as to maintain a constant population size. We have found that stronger selection \( (W_0 = 0.85) \) lengthens the expected takeoff time (Punishers are then more disadvantaged when rare), but does not otherwise materially affect our results.


10.3 Simulation Results

Unlike the analytical model, in which we formed groups using a relatedness coefficient, estimated from ethnographic hunter-gatherer populations (Table 7.1) $r = 0.07$ or $r = 0$, in the simulation the degree of genetic differentiation among groups is produced by the simulation itself, and is the result of the migration patterns and the distributions of group sizes across groups and over time, reflecting the parameters we have chosen and chance, given the random elements in mating and migration. Our strategy therefore is to adopt a parameter set that is both empirically plausible, and which in the simulation generates levels of genetic differentiation that are consistent with the estimates in Table 7.1, the mean of which is 0.08. Thus we ensure that we are not implicitly building into the simulation elevated levels of positive assortment that would favor the Punishment strategy.

As expected from our joint work Moreno Gamez and Wilkins discussed in Chapter 7, given variable group sizes estimated from ethnographic hunter-gatherer populations (Table 7.1) and other empirically plausible violations of the idealized assumptions of the equilibrium genetic differentiation model, the average group size of a single generation (30) and migration rate (38% of a group per generation) under which the simulation generates the observed mean value of the $F_{ST}$ (namely 0.077) are quite substantial. In the idealized model, these parameters would imply an equilibrium $F_{ST}$ of 0.021. The magnitude of the difference between the idealized and simulated $F_{ST}$ in this model is similar to that implied by the population genetic simulations with Moreno Gamez and Wilkins, confirming that the deviations from the ideal assumptions that our model produces are jointly of the same magnitude as those estimated empirically in the population genetics model.

Figure 10.5 shows a single simulation of the evolution of this model over 6000 generations (150,000 periods). The Punisher allele does not go to fixation because Nonpunishers have higher fitness than Punishers in groups in which the fraction of Punishers is far above the quorum threshold $\tau = 6$. The long-run average rate of non-cooperation after the emergence and proliferation of Punishers is 15%, one-fifth of which is due to behavioral error, and the remainder due to the fraction of Punishers falling below the quorum level in some groups through the migration, birth, and death processes.

To improve our understanding of the takeoff process and the subsequent evolutionary success of the Punisher strategy, we recall two terms, both derived from Price’s decomposition of evolutionary processes into additive
Figure 10.5. The Successful Invasion of Punishers and Stabilization of Cooperation. The timing of the takeoff to cooperation, which we define as 20% or less non-cooperation, depends on both the assumed rate of production of Punishers by the mutation process. In the simulations shown, the mutation rate is $\mu = 0.001$ and 3% of mutants are Punishers. The strength of the selection process is $1 - W_0 = 0.05$, and the quorum level is $\tau = 6$ for these simulations. The average $F_{ST}$ was 0.078 and the standard error of the $F_{ST}$ was 0.00641.

within-group and between-group selection effects presented in Chapter 6 (Price 1970). The first, $\beta_G$, is the effect of the fraction of a group that are Punishers on the group average expected fitness, while the second, $\beta_i$, is the effect of an individual’s own type (Punisher or not) on individual fitness. If $\beta_i$ is negative then punishing is altruistic, while $\beta_G$ may be positive if the presence of a large number of Punishers in group supports a high level of cooperation and hence high fitness for members of the group on average.

Figure 10.6 shows the average movement of these two coefficients in a process of taking off from zero to sustained cooperation. The take-off starts after about 23,500 periods. All curves are scaled down by a factor of 100 from periods 1 to 22350 to render them visible on the graph. The graph shows that before take-off punishment is not altruistic, as Punishers on average would do worse if they were to switch to Nonpunisher (the within-
group $\beta$ is greater than the across-group $b$). But after cooperation is consolidated, by period 25,000, Punishers are altruists: the within-group term is negative in most periods, indicating that Punishers do worse than Nonpunishers within groups on average. Also, after cooperation is consolidated, the across-group term is positive, indicating that on average, members of groups with a high frequency of Punishers have a fitness advantage over members of groups with few Punishers. The endogenously generated level of genetic differential among groups is such that most Punishers are in such groups and as a result their within-group disadvantage is offset, as described in equations 6.2.

The takeoff of Punishers occurs precisely when most Punishers, by the luck of the draw, are located in tipping groups, which are groups with just seven Punishers, so that if any one of them were to switch to Nonpunisher, the quota would not be met, punishing would cease and cooperation would collapse. This fact is illustrated by the simulation shown in Figure 10.7, which charts the probability that a Punisher is in a tipping group, which we term the ‘bunching statistic.’ For the simulation shown, this statistic shows...
a rapid, sustained increase during the takeoff period, after which it falls back to a low level when the frequency of Punishers is above 25%.

Figure 10.7. **Most Punishers are in tipping groups when punishing proliferates.** The Bunching Statistic is the probability that a Punisher is in a group where one individual moving from Nonpunisher to Punisher pushes the group from below to above the quorum level, or moving from Punisher to Nonpunishers pushes the group from above to below the quorum level.

Figure 10.8 illustrates the importance of quorum sensing for the operation of the model. In this figure, all parameters are at their benchmark levels except that we assume there is no quorum; i.e., Punishers always punish. Even stacking the deck in favor of Punishers by starting the simulation with a very high level of Punishers (75%), Punishers are driven by the invasion of Nonpunishers. After period 50,000, the rate of cooperation (not shown in the figure) is close to zero.

We have found that the Punisher type is not altruistic until cooperation is established. The reason for this is that when Punishers are rare, most successful punishing takes place in tipping groups, where a single Punisher defection leads to the collapse of cooperation. When Punishers are well-
established in the population, by contrast, most Punishers are in groups above the quorum threshold, where they are clearly altruists (any single Punisher would gain by switching).

Note that the same is true if the group is below the quorum: a Punisher would gain from switching to the Nonpunishing strategy, although he cannot know that beforehand. When the quorum is more than minimally met, a single Punisher would save both the signaling cost and the cost of participating in punishment by switching to the Nonpunishment strategy. It might therefore be thought that a new type would outcompete the Punishers. An Opportunist who acts exactly like a Punisher except that he defects in groups that exceed the quorum. However, the strategy of not punishing if the quorum is exceeded, which we call the Opportunist strategy, cannot displace Punishers in a cooperative equilibrium because when the ratio of Opportunists to Punishers in the population is high, and if cooperation is widespread, then most cooperative groups will exceed the quorum and will have a high ratio of Opportunists to Punishers. When the Opportunists switch to the non-punishing strategy, the quorum will no longer be met, and there will be no cooperation. Thus, when Opportunists are plentiful, they
will be predominantly in non-cooperating groups and their fitness will be low.

We show in Figure 10.9 that Opportunists cannot in fact become established in a population of Nonpunishers and Punishers. The figure shows that Opportunists can do quite well when cooperation is rare, but they eventually kill themselves off, allowing room for Punishers to invades, as they do around period 40,000 in the simulation.

Figure 10.9. The Dynamics of Opportunists in a Population of Punishers and Nonpunishers. Note that Opportunists do very well in a population of Nonpunishers, but eventually Punishers become established and Opportunists are driven from the population. The values shown represent a single run of the model with the baseline parameters as in the text.

One could imagine a complex Opportunist-type strategy in which when some but not all Opportunists are needed to support a quorum, they choose equitably among themselves who gets to defect and who remains cooperating with the Punishers. However, such a strategy would need reinforcement against preemptive Opportunists who simply refuse to punish, forcing the sophisticated Opportunists to share all the costs of punishing. This presents a serious second-order free-rider problem that cannot be handled using a
quorum-type strategy or any other we have been able to think of. Perhaps Nature has evolved a strategy of this sort in some creature, but it clearly does not exist in humans.

Additional details on the performance of the model are given in Appendix A11.

10.4 Why Quorum-based Punishment Succeeds

The key positive assortment that allows the stabilization of altruistic punishment at high levels in this model is not that the Punisher genotypes are more likely to be in groups with like genotypes. The success of Punishers is due to the fact that quorum sensing results in a genotype-phenotype correlation: when Punisher genotypes are common they are more likely to receive the benefits of cooperative actions by others, whether they are Punisher genotypes or not. As we saw in Chapter 6, (Queller 1992, Fletcher and Zwick 2006), such a correlation is necessary for natural selection to favor an altruistic genotype. Our simulations show that this situation is indeed the case. The correlation among groups between the phenotypic trait 'fraction cooperated' and the genetic trait 'fraction Punisher' becomes strongly positive, taking values between 0.5 and 0.55, when a state of high-level cooperation is attained. Experiments in dyads have suggested that punishment of defectors may lower group average payoffs (Dreber et al. 2008). This is not surprising, because in very small groups cooperation may be sustained by means of reciprocal altruist strategies such as tit-for-tat without resort to a punishment option. In the much larger groups studied here, the punishment option is essential to sustaining cooperation and high group average payoffs. It is also the case in our simulation that average fitness is much higher in groups with substantial fractions of Punishers than in groups where the frequency of Punishers is just above the threshold.

The critical genotype-phenotype correlation is enhanced by that fact that punishers coordinate their activities. Punishers do not punish unless their frequency in the group is sufficiently high to be cost-effective. This requires truthful communication among group members as to which members have shirked and which have not, as well as how many Punishers will share the costs of punishing. Thus, the effectiveness of altruistic punishment is predicated upon a broader social structure of mutual cooperation based on truthful communication and the widespread use of social sanctions.
If our speculation that strong reciprocity emerged through a modification of reciprocal altruist behaviors is correct, this provides another reason why strong reciprocity might be uniquely human, given that reciprocal altruism appears to be rare in other species (Stephens et al. 2002, Hammerstein 2003).

We have modeled a process of genetic transmission. But were the altruistic punishment trait subject to cultural transmission, as it certainly is in humans, then additional reasons for human exceptionality arise. The process of decay of a state in which virtually all group members are Punishers leading to a takeover of the group by Nonpunishers, one often observed in our simulations, could be slowed or even halted if cultural updating responded not only to payoff differences but also to the frequency of types in the group, so that more common behaviors were preferentially copied. Guzman et al. (2007) demonstrate that conformist cultural transmission of this type supports the evolution of altruistic punishment even for groups of enormous size (over 1,000 persons). They also demonstrate that a genetic predisposition to engage conformist updating could co-evolve with altruistic punishment in this environment.

10.5 Retribution and Social Order

Our model provides an explanation of the emergence and eventual stabilization of environments that may have supported high levels of cooperation among our ancestors living in mobile foraging bands during the late Pleistocene. We do not know that a human predisposition to strong reciprocity evolved as we have described. But our model and simulations suggest that it could have.

If cooperation did evolve by this route, it is not surprising that, as Cosmides and Tooby (1992) and others have shown, cheater detection is an advanced cognitive capacity of humans. First, humans are capable of inflicting punishment upon transgressors at very low cost to the punishers. While size, strength, and vigor generally determine the outcome of animal disputes, victory often involving great cost even to the winner, in human societies even a small number of attackers can defeat the most formidable single enemy at very low fitness cost to the attackers through the use of coordination, stealth and deadly weapons. Bingham (1999) has correctly stressed the importance of the superior abilities of humans in clubbing and throwing projectiles as compared with other primates, citing Goodall (1964), Plooij
(1978) on the relative advantage of humans, and Darlington (1975), Fifer (1987), and Isaac (1987) on the importance of these traits in human evolution. Calvin (1983) argues that humans are exceptional in possessing the same neural machinery for rapid manual-brachial movements that allow for precision stone-throwing.

Our model may also resolve one of the outstanding puzzles raised by behavioral experiments: the fact that while contributing to a public good and punishing those who fail to do so are both altruistic behaviors, subjects are considerably more avid about the latter than the former. As we saw in Chapter 5, the experiments of Ernst Fehr and his collaborators, as well as Fudenberg and Pathak (2009), show that inflicting punishment on norm violators is a strong human motive. The fact that some experimental subjects actively take pleasure in punishing transgressors, as indicated by their own accounts and by behavioral neuroscience experiments, could thus be the result of the evolutionary processes we have modeled.

Our model presents a sharp contrast to the repeated game approaches described in chapters 3 and 4. First, our model is consistent both with the empirical facts about forager societies outlined in Chapter 7. Second, unlike the fictive Rube Goldberg strategies invoked by the Folk Theorem, the main causal mechanism in our model, the altruistic punishment of norm violators, is widely observed, as we have seen in chapters 5 and 7.

Note that like the self-interest-based models that we criticized in Chapter 4, we have assumed that information is public. All members of a group see the same signals. Equivalently, if signals are not directly observed by all group members, they receive honest reports of the signals from those who did observe them. At the close of Chapter 4 we pointed out that this assumption is difficult to square with the assumption of self-interested behavior, since members will typically have something to gain by misrepresenting the actions taken by others. Modern large-scale societies, we observed, convert private to public information by judicial processes that took centuries to evolve and that presuppose that court officials, jury members, and law enforcement officers adhere to standards of professional conduct that preclude the unrestrained pursuit of self-interest. These models work only if they go beyond self-interest at a critical juncture. This is why the “self-interest with a long time horizon” explanation of seemingly generous acts fails.

The same reasoning applies to our model of the evolution of cooperation by means of the punishment of free-riders. Our model requires ethical and
other-regarding motives to render the public information assumption plausible. Smaller scale ancestral groups devised other ways to convert private to public information. Gossip, group discussions with all or most members present, and taking meals in public are examples. We think it likely that a norm of truth-telling developed *pari passu* with the punishment of freeriders. This does not mean that individuals were always truthful, but rather that there were effective means of punishing prevaricators and a reputation for honesty was a valuable resource in early human societies. We do not believe effective collective punishment could have evolved in the absence of a system information-sharing in which truth-telling is rewarded and lying punished.

We have focused entirely on punishing, saying nothing about why people would contribute to the public good other than that Punishers and Nonpunishers alike contribute when, in light of the likely punishment of shirkers, it is fitness maximizing to do so. But, were we to drop the fiction that fitness is explicitly maximized and introduce more empirically plausible proximate motives, it is something of a mystery how the kind of fitness maximizations we have assumed could occur. The benefits of shirking occur now, while the punishment, should it occur, happens later. Humans are too impatient to maximize their fitness and thus would not correctly weigh the future consequences of shirking even if they were correctly known.

We think that the evolution of the social emotions like shame may be explained not only as a support to the truth-telling essential to making accurate information public, but also in part by their ability enhance the present motivational salience of future punishments and thus to offset what would otherwise be a fitness reducing myopia of would be shirker’s. In the presence of altruistic punishment by one’s fellow group members, individuals with levels of impatience that lead them to shirk when it is not fitness maximizing to do so would improve their fitness were they to be motivated in part by feelings of shame and guilt. Christopher Boehm (2007) calls this process sanctioning selection. We will return to it when we address the social emotions in Chapter 12.

Thus individuals who developed the capacity to internalize group beneficial norms and to feel chastened when punished for violating these norms, as well as groups that devote their socialization practices to this end would appear to be evolutionarily favored. In the next chapter we show that this is indeed the case.
Socialization

The influence of society is what had roused in us the sentiments of sympathy and solidarity drawing us toward others; it is society which, fashioning us in its image, fills us with religious, political and moral beliefs that control our actions.

Durkheim, *Suicide* (1951[1897]) pp. 211–212.

Leges Sine Moribus Vanae
(Laws without Morals are Empty)

Horace, *Odes III.24*

In addition to trial and error experimentation, preferences are acquired by genetic inheritance (e.g., a taste for sweets) and by a learning process involving cultural transmission from our parents, others elders, and our peers (e.g., a taste for rice over potatoes). As we saw in Chapter 2, genetic and cultural transmission are in many ways similar, a fact has been exploited by the classic contributions to the modeling of cultural evolution by Cavalli-Sforza and Feldman (1981) and Boyd and Richerson (1985). The main similarity between the genetic and cultural processes that is exploited by these models is the fact that both social learning and genetic inheritance from parents can be represented as the replication of traits over time. Two additional similarities may be mentioned.

First, whether of cultural or genetic origin, the taste for sweets or rice activates the same reward-processing regions of the brain. The taste for sweets is certainly more universal among humans than is the taste for rice. But, there is no meaningful sense in which one can say that one is more deeply rooted or fundamental than the other. The genetically transmitted taste for sweets can easily be unlearned (a nauseating experience with sweet food overrides a genetic predisposition to like sweets, for instance). Similarly, culturally learned traits, such as the U.S. Southern culture of honor (Nisbett and Cohen 1996), have physiological correlates, such as elevated
testosterone when insulted, much as physical danger elevates adrenaline in virtually all humans.

Second, those who are relatively successful in acquiring material resources tend to produce more copies of their traits in the next generation, whether the process works through their differential success in producing offspring who survive to reproductive age or because of their greater command of resources, more elevated social status, or other reasons for their greater likelihood of being copied as cultural models. In the previous chapters we have specified evolutionary processes in which the frequency of a behavioral type in the population increases if its expected payoff exceeds the average. These so-called “payoff-monotone models” provide a challenging, if highly simplified, way of posing the puzzle we are addressing, namely, the evolution of preferences that induce people to act in ways that reduce their payoffs by comparison to what they would get if they acted in some other manner. In the models proposed in chapters 8, 9, and 10, altruistic traits may overcome their within-group payoff disadvantages first, because of the superior payoffs enjoyed by members of groups in which there are many altruists and second, because groups devise, and culturally transmit over generations, the institutions that mitigate the within-group selection pressures tending to eliminate altruists.

Cultural transmission provides an additional way that the payoff disadvantages of altruistic social preferences might be overcome. In his book Sick Societies, Robert Edgerton (1992) catalogues dozens of examples of culture overriding fitness, all, as the title suggests, with unpleasant consequences.

Pre-industrial cities provide an example (Knauft 1989). Prior to modern medicine the city was a cultural success, recruiting steady streams of migrants to forsake the countryside for urban living. But, it was a biological failure, typically not reproducing its own population even among the social elites residing therein. A second example is the demographic transition, whereby the culturally-transmitted preference for smaller families proliferated in many populations despite having apparently reduced fitness (Zei and Cavalli-Sforza 1977, Kaplan et al. 1995, Ihara and Feldman 2004).

But, if cultural transmission can induce people to limit their fitness by having small families, or to choose a lethal residential environment, it certainly might also overcome the payoff disadvantages associated with altruistic social preferences. It is this possibility that we explore here. The puzzle, of course, is to explain why humans or any other animal would ever develop
the capacity to override fitness concerns, for that capacity itself would seem
to be doomed by natural selection.

11.1 Cultural Transmission and Socialization

Cultural transmission overrides fitness when it causes people to want to do
things that result in their having fewer surviving offspring. Thus our expla-
nation will involve the proximate causes of behavior, that is to say, prefer-
ences. Here, and in the next chapter as well, we depart from the framework
of the previous three chapters, which have focused entirely on fitness and
behavior without exploring the question of motivation. It is not difficult, of
course, to associate proximate motives with the kinds of behaviors that we
have shown may evolve. Ethically motivated outrage, what Robert Trivers
called “moralistic aggression,” is a plausible motivation for the strong re-
ciprocators’ punishment of defectors in Chapter 10, and group loyalty and
group hostility could provide the psychological basis for the behaviors
studied in Chapters 8 and 9. Our models show that these and other prefer-
ences motivating the behaviors in question could have evolved by a fitness
based evolutionary process. Here we seek to understand how altruistic pref-
ferences might evolve under the influence of cultural transmission.

We will take account of two facts. First, the phenotypic expression of an
individual’s genetic inheritance depends on a developmental process that
is plastic and open-ended. One expression of this fact is that while human
ancestral groups are similar genetically (Feldman et al. 2003), they differ
in important ways in behaviors. We surveyed some of our experimental
evidence for this behavioral variability in Chapter 5. This developmental
plasticity explains why humans are among the most ubiquitous of species,
capable of making a living and surviving in virtually all of the world’s en-
vironments.

Second, this developmental process is deliberately structured, by elders,
teachers, political leaders, and religious figures, to foster certain kinds of
development and to thwart others. In many of Edgerton’s sick societies,
the socialization processes affecting development results in proximate mo-
tives leading people to engage in such lethal practices as cigarette smoking
or, in the highlands of New Guinea, consuming the brains of deceased rela-
In both cases individuals contract a terminal illness with high probability.
But, in most societies, socialization stresses not only the desirability of be-
haviors that contribute to one’s own well-being, such as moderation, planning ahead, and personal hygiene, but also those that benefit others, such as the altruistic social preferences and character virtues we have identified as common among humans.

In this chapter we analyze the process by which social norms become internalized, that is, taken on as preferences to be sought in their own right rather than constraints on behavior or instrumental means to other ends. Internalization is thus an aspect of cultural transmission that affects preferences rather than beliefs and capacities. The idea of internalized norms is captured in a passage attributed to Abraham Lincoln: “when I do good, I feel good. When I do bad, I feel bad. That is my religion.”

Much of the content of cultural transmission can be modeled as information transfer. Members of a group, most often as children, are taught ‘how to’ accomplish particular ends such as acquiring and preparing food or performing music, as in the study of the Central African Aka by Barry Hewlett (1986). We focus instead on the process by which a society’s ‘oughts’ become its members ‘wants,’ thereby narrowing the hiatus between what Jeremy Bentham famously termed people’s ‘dutys’ and their ‘interests.’ As a result, we draw upon studies of how values, rather than factual information, are transmitted, such as generosity among the Inuit (Guemple 1988), social solidarity among children on Israeli kibbutzim (Bronfenbrenner 1969), and the control of hostility among children in cross-cultural perspective (Whiting and Whiting 1975). We refer to these ‘ought’ rules of behavior as norms and when they are internalized, as preferences.

Though drawing on a somewhat different mix of social institutions for its accomplishment, the internalization of norms has enough in common with other aspects of cultural transmission that we can draw upon the models of Boyd and Richerson and of Cavalli-Sforza and Feldman. We posit three influences on the cultural transmission of preferences and model how they may interact so as to favor the evolution of other-regarding and ethical preferences. First, we model the vertical transmission of traits from parents to offspring. Parental traits that are associated with greater fitness will evolve for the same reason that genes that confer greater fitness enjoy supra-average survival rates. The second is oblique transmission to the young from non-parental members of the parents’ generation in the myriad of personal interactions with neighbors, teachers, and spiritual leaders by which the young are socialized to internalize particular norms (Cavalli-Sforza and Feldman 1981). Third is payoff-based social learning according
to which periodically, over the life course, people compare their behaviors with the behaviors of other individuals, and tend to adopt behaviors of others who appear to be doing relatively well. We take account of the effect of payoffs on the adoption of norms in order to counter the oversocialized concept of the individual according to which socialization simply implants norms in a passive and uncritical target (Wrong 1961, Gintis 1975).

Following Boyd and Richerson (1985), oblique transmission may be conformist, the young tending to adopt the behaviors most common in the parental generation, irrespective of their payoffs. In this case resulting dynamic will not be monotonic in either fitness or well-being. If virtually all of the population is altruistic, conformist cultural transmission might overcome the payoff disadvantage suffered by the altruists and allow their persistence in a population. Conformism may also stabilize payoff-reducing behaviors that yield no benefit to others, such as smoking. Indeed, this is the most parsimonious explanation of the long term persistence of many of the dysfunctional behaviors documented by Edgerton. Conformism may thus contribute to large between-group differences in behavior, with selection against low payoff behaviors within groups being weak or absent. In the presence of strong conformism, weak group selection may be sufficient to stabilize altruistic preferences. Conformist cultural transmission is unquestionably an important influence on the evolution of preferences.

Conformist cultural transmission may arise for a variety of reasons, ranging from an evolved social learning strategy in which individuals regard the population frequency of a trait as a measure of its desirability, all the way to population-level institutional arrangements for the deliberate socialization of the young, in which the content reflects which types are prevalent in the population. We stress the latter for empirical reasons: most societies devote substantial time and resources to deliberately socializing the young to act in ways that are beneficial to others, and an adequate explanation of social preferences needs to take account of this fact.

Why should the norms that are internalized be altruistic? Linnda Caporeal Caporael et al. (1989) and Herbert Simon Simon (1990) proposed that altruism might proliferate in a population because it is an inseparable part of a ensemble of culturally transmitted norms that is, on balance, individually advantageous. Simon termed the capacity to internalize such an ensemble of social norms docility (literally, ‘teachability’) and explained the evolution of altruistic behaviors as a consequence of the fact that the norms motivating them are linked to other norms that benefit the individual
sufficiently to offset the individual costs of altruism. Altruism in this case proliferates in the same way that a genetically transmitted disadvantageous trait may evolve if it is pleiotropically linked to other, advantageous traits and thus may hitchhike on their success.

We wish to explore this reasoning and address two aspects in which it is incomplete. First, one needs to address the puzzle of how the capacity to internalize norms evolves (§11.2). Second, we would like to relax the *ad hoc* “pleiotropic analogy” whereby individually costly altruism and individually beneficial other norms are inseparable. In the model we present in §11.3 and §11.4, as in the Simon model, altruism will ‘hitchhike’ on the beneficial aspects of internalizing other norms. But here the hitchhiker is not forced upon the driver, and rather is picked up voluntarily, the endogenous result of a gene-culture evolutionary dynamic. In §11.5 we incorporate the insight that the higher the cost of holding an altruistic norm, the more likely are individuals to abandon that norm. We accomplish this by adding a payoff-based updating process to our dynamical system. In §11.6 we show that even when individually beneficial and individually costly altruistic behaviors are not linked other than by the fact that both may be internalized, altruism may successfully hitchhike under plausible parameter values. This is why, despite the evidence provided by Edgerton, institutions of socialization tend to favor prosocial preferences.

In these models the existence of an internalized norm that enhances group average fitness is critical to the evolution of the capacity for internalization. But developing the capacity to internalize norms is costly to the individual, and sustaining the institutions whereby internalization takes place is costly to society. Why would evolution favor bearing these costs rather than relying on genetic transmission to sustain individually beneficial norms? The answer we propose in §11.7 is an application of the reasoning of Boyd and Richerson (2000), extending the explanation give in Chapter 2. The cultural transmission of norms allowed humans exceptionally among animals to adapt flexibly to rapidly changing circumstances and to modify the results of individual fitness maximization where these are not beneficial on average to members of a group.

### 11.2 When is Virtue its Own Reward?

Consider a group in which members can either adopt, or not, a certain cultural norm A. We shall call those who adopt the norm A altruists, or A-types,
and we call those who do not adopt norm A self-regarding types, or S-types. Altruism is costly, in that S-types have fitness 1, as compared with A-types, who have fitness $1 - s$, where $0 < s < 1$ is a viability loss. We assume in each generation that individuals pair off randomly, mate, and have offspring in proportion to their fitness, after which they die. Families pass on their cultural norms to their offspring, so offspring of AA parents are A-types, offspring of SS parents are S-types, and half of the offspring of AS-families (which are the same as SA-families) are A-types, the other half S-types (we call this vertical transmission). We also assume that the S-type offspring of AS- and SS-families are susceptible to influence by socialization institutions promoting altruistic norms, a fraction of such offspring becoming A-types (oblique transmission).

We find that the change in the fraction of A-types in the next generation is given by the familiar replicator equation (see Appendix A2):

$$
\Delta f_A = f_A (1 - f_A) \frac{\gamma - s}{1 - sf_A},
$$

where $f_A$ is the frequency of A’s in the population, $1 - sf_A$ is the average payoff in the population, $f_A (1 - f_A)$ is a measure of the frequency of AS pairings in the population, and $\gamma$ is the rate of oblique transmission, so $\gamma - s$ gives the selective advantage (or disadvantage) of the A’s over the S’s when account is taken of both oblique and vertical transmission. Equation (11.1) illustrates the tension between the differential fitness effects on the evolution of $f_A$ captured by $s$ that work against the evolution of the A’s and the effects of oblique transmission captured by $\gamma$, which tend to counteract the selection against A-types. Equation (11.1) shows that when $s = \gamma$, these two effects are exactly offsetting, and the population frequency of A-types will be stationary ($\Delta f_A = 0$).

Payoff-based updating then occurs. Each group member $i$ observes the fitness and the type of a randomly chosen other member $j$, and changes to $j$’s type if $j$’s fitness is higher. However, information concerning the difference in fitnesses of the two strategies is imperfect, and individuals’ preference functions do not perfectly track fitness, so it is reasonable to assume that the larger the difference in the payoffs, the more likely the individual is to perceive it, and change. Specifically, we assume the probability that an A individual will shift to S is proportional to the fitness difference of the two types, so $p = \eta s$ for some imitation rate $\eta > 0$. The term $\eta$ represents the power of payoff differences to induce changes in type, and this, naturally will play a big role in our account.
The expected fraction \( f_A' \) of the A population after the above shifts is the fraction before updating \( f_A \), minus those A’s who switched to S, the latter being the A’s who were paired with a S, who constitute a fraction \( f_A(1 - f_A) \) of the population, multiplied by the probability of a switch taking place in these cases. Thus we have

\[
f_A' = f_A - \eta s f_A(1 - f_A).
\] (11.2)

We now combine these three sources of change in the fraction of A-types, adding the changes described in equation 11.2 to those already shown in 11.1, giving

\[
\Delta f_A = f_A(1 - f_A) \frac{\gamma - s}{1 - s f_A} - \eta f_A(1 - f_A) s
\] (11.3)

The second term on the right hand side represents the influence of payoff-based updating, reducing the frequency of the altruistic norm, in comparison with the vertical and oblique cultural transmission mechanisms, represented by the first term, which may favor this norm or not, depending on whether \( \gamma > s \) or \( \gamma < s \).

Not surprisingly, the higher the personal cost of altruistic behavior, the more stringent the conditions under which altruism will emerge, illustrating the tension between socialization institutions and the psychological mechanism of norm internalization on the one hand, and payoff-based updating that induces individuals to shift to higher payoff behaviors, whatever the effect of these behaviors on others, and on society as a whole, on the other hand. This tension is revealed by the conditions under which the altruistic equilibrium is globally stable, meaning that starting from any of the possible states of the population, the population dynamic will lead move to the all-S equilibrium, if the strength of payoff-based updating \( \eta \) is less than the difference in the size of the oblique transmission and the fitness cost of altruism, normalized by the latter:

\[
\eta < \frac{\gamma - s}{s},
\] (11.4)

However, if

\[
\frac{\gamma - s}{s} < \eta < \frac{\gamma - s}{s(1 - s)},
\] (11.5)

both the S-type and the altruistic equilibria are locally stable, and the basin of attraction of the altruistic equilibrium, that is, the states from which the
dynamic will converge to the all $A$ equilibrium, shrinks as $\eta$ increases. Finally, if
\[ \eta > \frac{\gamma - s}{s(1 - s)}, \]
the all-$S$ equilibrium is globally stable.

Thus if the internalization of norms accomplished by the society’s socialization processes ($\gamma$) is sufficiently strong relative to the strength of payoff-based updating ($\eta$) and the cost of altruism ($s$), the altruistic equilibrium may be stable. In effect, there is a net flow into altruism at rate $\gamma$, the rate of oblique transmission, a net flow out of altruism due to its fitness cost $s$, and another flow out because individuals switch from altruistic to self-regarding behavior by copying the more successful self-regarding individuals, at rate $\eta$. When the net balance favors a positive flow into altruism, i.e., when $\gamma > s + s\eta(1 - s)$, the altruistic equilibrium is at least locally stable.

### 11.3 The Coevolution of Genes and Culture in Norm Internalization

But why would people, or any animal, internalize norms if taking a norm on board leads one to act in ways that reduce fitness? We will answer this in two steps. In this section we will explain why the capacity to internalize fitness enhancing norms, ones that correct for human impatience or weakness of will for example, might evolve even if the capacity to internalize is costly. In the next section we show that when the capacity to internalize a norm has evolved and societies have developed socialization practices to do this, people will be susceptible to internalizing norms that also reduce fitness. This is what we mean when we say that altruism can *hitchhike* on a process of norm internalization that has evolved due to the existence of an individually fitness enhancing social norm.

Our model of the internalization process solves a serious problem with the Caporeal-Simon notion that fitness-enhancing and altruistic cultural traits come as an inseparable bundle. The problem is to explain why the bundle remains inseparable. A self-regarding agent might easily adopt the personally fitness-enhancing parts of the bundle and reject the others. In our model the capacity to internalize norms is genetically predisposed, so an individual who has this genetic predisposition, having internalized a set of
socially-provided norms, incorporates these norms directly in his preference function, and hence has no interest in rejecting the purely altruistic members of the set, unless, of course, the cost of compliance is excessive, a contingency we investigate below.

Here we assume that cultural traits are acquired from parents or through oblique transmission, but that payoff based switching of traits, as modeled in the previous section, does not occur. We introduce payoff based cultural updating in this gene-culture coevolutionary model in §11.4.

To simplify the analysis we assume that there is one genetic locus that controls the capacity to internalize norms, and that norm internalization is the expression of a single allele, which we will call the ‘internalization allele’. This is considerable simplification, but a more complex treatment would not provide any additional illumination of the questions under investigation. We will assume that each individual has only one copy at this locus (i.e., genetics are haploid), which can be inherited with equal probability from either parent (an alternative diploid model, in which each locus has two alleles at each locus, has almost the same properties as the haploid model, but is much more complicated, and is developed in full in Gintis 2003a). Individuals without the proper allele cannot internalize norms, whereas individuals with the proper allele are capable of internalization, but whether or not they internalize a norm depends on costs and benefits, as well as the individual’s personal history. In this section we assume that an internal norm is fitness enhancing and we derive the conditions under which the allele for internalization of norms is globally stable. Suppose the norm in question is C (Cleanliness, for instance), which confers fitness $1 + f > 1$, while the normless phenotype, denoted by D (Dirty, perhaps), has baseline fitness 1. There is a genetic locus with two alleles, $a$ and $b$. Allele $a$ permits the internalization of norms, whereas $b$ does not. We assume that possessing $a$ imposes a fitness cost $u$, with $0 < u < 1$, on the grounds that there are costly physiological and cognitive prerequisites for the capacity to internalize norms. We assume $(1 + f)(1 - u) > 1$, so the cost of the internalization allele is more than offset by the benefit of the norm C. An individual is now characterized not only by his genes, but his phenotype (whether he is a $a$ or a $b$). There are thus three “phenogenotypes,” whose fitnesses are shown in Table 11.1.

The rules of gene-culture transmission are as follows. If a familial phenogenotype is $xyXY$, where $x$ and $y$ can be either $a$ or $b$, and X and Y can be either C or D, an offspring is equally likely to inherit $x$ or $y$. An
Table 11.1. **Fitnesses of the Three Phenogenotypes.** Here $u$ is the fitness cost of possessing the internalization allele, and $f$ is the fitness value of possessing the norm $C$. Note that $bC$ cannot occur because an individual must have $a$ to be able to internalize $C$.

<table>
<thead>
<tr>
<th>Phenogenotype</th>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>$aC$</td>
<td>$(1 - u)(1 + f)$</td>
</tr>
<tr>
<td>$aD$</td>
<td>$1 - u$</td>
</tr>
<tr>
<td>$bD$</td>
<td>$1$</td>
</tr>
</tbody>
</table>

Table 11.2. **Phenotypic Inheritance is Controlled by Genotype.** Note that $bCC$ and $bCD$ are not listed. This is because $bC$ cannot occur, because an individual must have the $a$ allele to internalize the $C$ norm. Note that $\beta \in [0, 1]$ measures the strength of the cultural transmission of $C$.

<table>
<thead>
<tr>
<th>Familial Type</th>
<th>Offspring Phenogenotypic Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$aaCC$</td>
<td>$aC$</td>
</tr>
<tr>
<td>$aaCD$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>$aaDD$</td>
<td></td>
</tr>
<tr>
<td>$abCD$</td>
<td>$\beta/2$</td>
</tr>
<tr>
<td>$abDD$</td>
<td>$1/2$</td>
</tr>
<tr>
<td>$bbDD$</td>
<td></td>
</tr>
</tbody>
</table>
three familial phenotypes, CC, CD, and DD, and nine familial phenogenotypes, of which only six can occur (because a parent of genotype b must have the D phenotype). The frequencies of the offspring of different familial phenogenotypes are as shown in Table 11.3, where $P(i)$ represents the frequency of parental phenogenotype $i = aC, aD, bD$. For example, the $aaCD$ phenogenotype can occur in two ways: father $aC$ and mother $aD$, or vice-versa. The probability of each occurrence is $P(aC)/P(aD)$. The fitness of this phenogenotype is $(1-u)^2(1+f)$ because both parents have the $a$ allele at fitness cost $u$, and one has the C trait, at fitness gain $f$. The proportionate share of the offspring of this phenogenotype is thus as given in the second line of Table 11.3.

<table>
<thead>
<tr>
<th>Phenogenotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$aaCC$</td>
<td>$P(aC)^2(1-u)^2(1+f)^2\beta_o^2$</td>
</tr>
<tr>
<td>$aaCD$</td>
<td>$2P(aC)P(aD)(1-u)^2(1+f)\beta_o^2$</td>
</tr>
<tr>
<td>$aaDD$</td>
<td>$P(aD)^2(1-u)^2\beta_o^2$</td>
</tr>
<tr>
<td>$abCD$</td>
<td>$2P(aC)P(bD)(1-u)(1+f)\beta_o^2$</td>
</tr>
<tr>
<td>$abDD$</td>
<td>$2P(aD)P(bD)(1-u)\beta_o^2$</td>
</tr>
<tr>
<td>$bbDD$</td>
<td>$P(bD)^2\beta_o^2$</td>
</tr>
</tbody>
</table>

Table 11.3. Frequencies of Phenogenotypes. Here, $\beta_o$ is baseline fitness, and is chosen so the sum of the frequencies is unity. Note that $bCC$ and $bCD$ are not listed, because $bC$ cannot occur.

Equilibrium occurs when each phenogenotype has constant generation to generation frequency. In this case, we need consider only two of the phenogenotypes, say $aC$ and $aD$, because $bC$ cannot occur, and since the probabilities must add up to unity, we have $P(bD) = 1 - P(aC) - P(aD)$. This system has three equilibria, in which the whole population bears a single phenogenotype. These are $aC$, in which all individuals internalize the fitness enhancing norm, $aD$, in which the internalization allele is present but the phenotype $C$ is absent, and $bD$, in which neither the internalization allele nor the norm is present.

Elsewhere (Gintis 2003b) we have proven the following assertions concerning the stability of the various equilibria of this system. The $aD$ equilibrium is unstable, while the $aC$ equilibrium is locally stable. The unnormed equilibrium $bD$ is locally stable if $(1-u)(1+f) < 2$, and unstable when the opposite inequality holds. There are two conditions that render the $bD$
equilibrium unstable, in which case $aC$ will be globally stable. The first is that $(1 - u)(1 + f) > 2$. The second condition is that the cultural bias transmission coefficient $\beta$ is sufficiently greater than $1/2$ and $(1 - u)(1 + f) > 1$. We consider the condition $(1 - u)(1 + f) > 2$ rather implausible because it requires that $f > 1$, whereas positive fitness coefficients are rarely so large. However, the biased transmission condition, $\beta > 1/2$, is quite plausible, because it may take only one parent to instill a norm in all offspring with high probability (“Mom taught me to be clean. Dad was always a slob”).

11.4 Altruism as Hitchhiker

We now add a second phenotypic trait with two variants. Internal norm $A$ is altruistic in the sense that its expression benefits the group, but imposes fitness loss $s$, with $0 < s < 1$ on those who adopt it. The normless state, $S$, is neutral, imposing no fitness loss on those who adopt it, and also no gain or loss to other members of the social group. An individual phenotype is then one of $SD$ (internalizes neither norm), $SC$ (internalizes only the fitness-enhancing norm), $AD$ (internalizes only the altruistic norm), and $AC$ (internalizes both the fitness-enhancing and altruistic norm).

We assume $A$ has the same cultural transmission rules as $C$: $a$-individuals inherit their phenotypes from their parents, while $b$-individuals always adopt the normless phenotype $SD$. In addition, there is oblique transmission, as before. There are now two genotypes and four phenotypes, giving rise to five phenogenotypes that can occur, which we denote by $aAC$, $aAD,aSC$, $aSD$, and $bSD$, and the three that cannot occur because $b$-individuals must be normless; i.e. $SD$. These three are $bAC$, $bAD$, and $bSC$. We represent the frequency of phenogenotype $i$ by $P(i)$, for $i = aAC, \ldots, SD$.

As before, families are formed by random pairing and the offspring genotype obeys Mendelian segregation. We assume also that only the phenotypic traits of parents, and not which particular parent expresses them, are relevant to the transmission process. Therefore there are nine family phenotypes, which can be written as $AACC$, $AACD$, $AADD$, $ASC$, $ASC$, $ASCD$, $ASDD$, $SSC$, $SSC$, and $SSDD$. It follows that there are 27 familial phenogenotypes, which we can write as $aaAACC, \ldots, bbSSDD$, only 14 of which can occur. We write the frequency of familial phenogenotype $j$ as $P(j)$, and we assume the population is sufficiently large that we can ignore random genetic drift. For instance, $aaAACC$ represents the case where both parents
Chapter 11

\[
P(aaAACC): \ P(aaAC)^2(1-u)^2(1+f)^2(1-s)^2\beta_o^2,
\]

\[
P(aaAADC): \ 2P(aaAC)P(aAD)(1-u)^2(1-s)^2(1+f)\beta_o^2,
\]

\[
P(abASCD): \ 2P(2aaAC)P(bSD)(1-u)(1+f)(1-s)\beta_o^2,
\]

\[
P(bbSSDD): \ P(bSD)^2\beta_o^2.
\]

Table 11.4. Here, \(\beta_o\) is baseline fitness, and is chosen so the sum of the frequencies is unity. To understand this calculation, consider, for instance the \(abASCD\) phenogenotype. This can arise in two ways: (1) \(aAC\) mother and \(bSD\) father or (2) \(bSD\) mother and \(aAC\) father. In both cases, one parent came from a pool with fitness \((1-s)(1+f)(1-u)\) and the other with fitness 1.

have the internalization allele \(a\), and both parents internalize the altruistic norm \(A\) and the fitness-enhancing norms \(C\). Similarly, \(aaAADC\) represents the case where both parents have the internalization allele \(a\), and both parents internalize the altruistic norm \(A\), but only one internalizes the fitness-enhancing norm \(C\). Finally, \(abASCD\) represents the case one parent carries the internalization norm and the other does not, the former internalizing both the altruistic norm \(A\), and the fitness-enhancing norm \(C\). We write the frequency of familial phenogenotype \(j\) as \(P(j)\), and we assume the population is sufficiently large that we can ignore random drift. For illustrative purposes, a few of the phenogenotypic frequencies are shown in Table 11.4.

The rules of cultural transmission are as before. If familial phenogenotype is \(xyXYZW\), where \(x\) and \(y\) are either \(a\) or \(b\), \(X\) and \(Y\) are either \(A\) or \(S\), and \(Z\) and \(W\) are either \(C\) or \(D\), an offspring is equally likely to inherit \(x\) or \(y\). An \(a\) offspring is equally likely to inherit \(X\) or \(Y\), and equally likely to inherit \(Z\) or \(W\). Offspring of genotype \(b\) always have the normless phenotype \(SD\). Oblique cultural transmission occurs when an \(a\)-individual with \(S\) phenotype, genetically capable of internalizing but culturally selfish, adopts the \(A\) phenotype in response not to parental socialization but to learning from other \(A\)-types in the population. This occurs more frequently the more \(A\)-types there are in the population (\(p_A\)) and the more effective are the society’s institutions (deliberate or otherwise) for oblique transmission (\(\gamma\)), each \(aS\) individual switching at the rate \(\gamma p_A\), so that the gain in \(A\) phenotypes by this mechanism is \(\gamma p_A p_S\), where \(p_S\) is the frequency of \(aS\) types in the population. Note that oblique transmission in this model is asymmetric: if there are \(A\) types in the population, \(S\)-types may learn to become \(A\) types, not the other way around, even if the population is predominantly of the \(S\)-type.
We assume both genotypic and phenotypic fitness, as well as their interactions, are multiplicative. Thus, for instance, an \( aAC \) individual incurs a fitness cost \( u \) from the capacity to internalize, a fitness gain of \( f \) from holding norm \( C \), and a fitness loss \( s \) from holding the \( A \) norm. The individual’s resulting fitness is then \( (1 - u)(1 + f)(1 - s) \). In the absence of positive assortment, \( (1 - u)(1 + f)(1 - s) > 1 \) is a necessary condition for the altruistic trait to evolve, so we assume this inequality holds; i.e., the direct individual fitness benefit due to having phenotype \( C \) must be sufficient to offset both the cost of having the internalization allele and the cost of altruism. The fitness of the phenogenotypes that can occur with positive frequency are as shown in Table 11.5.

<table>
<thead>
<tr>
<th>Individual Phenogenotype</th>
<th>Individual Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>( aAC )</td>
<td>( (1 - u)(1 - s)(1 + f) )</td>
</tr>
<tr>
<td>( aAD )</td>
<td>( (1 - u)(1 - s) )</td>
</tr>
<tr>
<td>( aSC )</td>
<td>( (1 - u)(1 + f) )</td>
</tr>
<tr>
<td>( aSD )</td>
<td>( (1 - u) )</td>
</tr>
<tr>
<td>( bSD )</td>
<td>( 1 )</td>
</tr>
</tbody>
</table>

Table 11.5. **Fitnesses of Five Phenogenotypes.**

The fitness of these phenotypes, along with the rules of genetic and cultural transmission given above, allow us to determine for any combination of frequencies of the phenogenotypes in Table 11.5 the change in frequencies that will occur. The population is in equilibrium when the frequency of each phenogenotype is constant from generation to generation. We can determine the possible population equilibria using four equations, one each for the constancy of frequency of \( aAC \), \( aAD \), \( aSC \), and \( aSD \), the frequency of \( bSD \) being one minus the sum of the other frequencies. These equations show that there are five equilibria, in which the whole population bears a single phenogenotype. These are \( aAC \), in which all individuals internalize both the altruistic and fitness enhancing norms, \( aAD \), in which only the altruistic norm is internalized, \( aSC \), in which only the fitness-enhancing norm is internalized, \( aSD \), in which individuals carry the allele for internalization of norms, but no norms are in fact internalized, and \( bSD \), in which internalization is absent, and neither altruistic nor fitness-enhancing norms are transmitted from parents to offspring. But the \( aAD \) and the \( aSD \) equilibria
are unstable, and hence will not survive an evolutionary process, so we can ignore them.

The analysis of the stability of the remaining equilibria is given in Gintis (2003a). Stability holds when $s < \gamma$. This inequality expresses the key condition that altruism cannot be evolutionarily stable unless the effectiveness of oblique transmission is sufficient to overcome the fitness cost of altruism. Groups with high levels of altruism solve the problem of rendering altruism fitness-enhancing by increasing the effectiveness of oblique transmission so that the new converts to altruistic norms compensate for the lower fitness of A-types.

It is no surprise, therefore, that the $a_{SC}$ equilibrium, in which internalization is possible but the altruistic norm is not internalized, is stable when $\gamma < s$, and unstable when the opposite inequality holds. This reinforces the interpretation presented in the previous paragraph. Moreover, as in the single phenotype case, $b_{SD}$ is unstable if $(1 - u)(1 + f) > 2$, which is highly unlikely as we explained above. There are two reasons why the equilibria $a_{SC}$, $a_{AC}$ and $b_{SD}$, all homogeneous populations with a single type, are stable. First, there are positive feedbacks in the oblique transmission process by which altruists are socialized, such that it is totally inoperative when $p_A = 0$ and may be powerful enough to offset the fitness disadvantages of the altruists when $p_A = 1$. This explains why a stable equilibrium population is either all S or all A. Second the $b_{SD}$ (“no internalization, no norms”) equilibrium is stable is that $a$ and $C$ are complements, meaning that in the absence of $C$, $a$ cannot proliferate when rare, and conversely. We have not determined if stable mixed strategy equilibria exist, but for the above reasons we doubt that they could.

This analysis shows that if $s < \gamma$, the altruistic phenotype A coexists in a stable equilibrium with the fitness-enhancing phenotype C. We say that A hitchhikes on C in the sense that the fitness value of C renders the internalization allele $a$ evolutionarily viable, and once this allele occurs in high frequency, the altruistic phenotype A is evolutionarily viable because its fitness cost $s$ is less than the oblique transmission effect $\gamma$, which favors A.

### 11.5 How Altruism Survives Payoff-based Updating

But we stacked the deck to show that equilibria with a genetically-determined capacity for internalization and a culturally-transmitted altru-
istic norm are possible. To simplify the gene-culture interaction we did not include an obvious challenge to the altruistic trait: when people update their behaviors they not only do so under the influence of schools, elders and the other bearers of oblique transmission, they also pay attention to the payoffs that they and people of different types are receiving, and this must disadvantage the altruists. We now add the payoff-based updating dynamic developed in §11.2 to our gene-culture model, thus allowing individuals to shift from lower to higher payoff strategies, and we show that the result is similar to that of the model developed without genetics in §11.2. In the current context, there are four phenotypes only $a$-individuals will copy another phenotype, because only such types are capable of internalizing a norm, and noninternalizers will not desire to mimic internalizers.

We assume an $a$-individual with phenotype XY meets an individual of type WZ with probability $p_{WZ}$, where $p_{WZ}$ is the fraction of the population with phenotype WZ, and in this case switches to WZ with probability $\eta$ if that type has higher fitness than XY. Thus, as in §11.2, the parameter $\eta$ is the measure of the strength of the tendency to shift to high-payoff phenotypes.

Adding payoff-based updating does not change the single phenogenotype equilibria, because all equilibria consist of a single phenogenotype, so in equilibrium, an individual can never meet a distinct phenotype to which he might switch. We find that the $aAD$ and $aSD$ equilibria remain unstable, and payoff-based updating does not affect the conditions for stability of the unnormed equilibrium $bSD$. The condition $\gamma > s$ for stability of the altruistic equilibrium $aAC$ now becomes

\[
\eta < \frac{\gamma - s}{1 - \gamma} \left( \frac{1}{s} - 1 \right). \tag{11.7}
\]

Note the similarity to the altruistic equilibrium conditions (11.4–11.6) in the model without the explicit modeling of genetics. We conclude that a sufficiently strong payoff-based updating process can undermine the stability of the $aAC$ equilibrium. The condition $s > \gamma$ for stability of the nonaltruism internalization equilibrium $aSC$ when payoff-based updating is included now becomes

\[
\eta > \frac{\gamma - s}{s(1 + \gamma - s)},
\]

and this equilibrium is unstable when the reverse inequality holds. Thus in this case, $s > \gamma$ continues to ensure that $aSC$ is stable, but now for sufficiently large $\eta$, this equilibrium is stable even when $\gamma > s$. 
In sum, adding payoff-based updating changes the stability properties of
the model in only one important way: a sufficiently strong payoff-based
updating process can render the nonaltruistic yet internalized equilibrium
aSC, rather than the altruistic equilibrium, aAC, stable. The intuition here
is that altruism imposes a fitness cost $s$ leading individuals to abandon altrui-
sm. The greater the rate at which this occurs, the larger must be the oblique
socialization force $\gamma$ that replenishes the stock of A-types in the group.

We may summarize the model as follows. In the presence of C norms,
internalization is individually fitness-enhancing, i.e., $(1 - s)(1 + f) > 1$, and hence the $a$ allele undergoes positive selection to fixation (all $a$).
Altruism is costly, i.e., $u > 0$, but by assumption, the net fitness of A-types,
$(1 - s)(1 + f)(1 - u)$ remains greater than that of non-internalizers, who
have fitness unity. Because $u > 0$, internalizers tend to abandon altruism
at a positive rate, both because they are at a reproductive disadvantage and
they abandon altruism for the selfish alternative at a positive rate, so only
an additional force will allow altruism to evolve. The major such forces is
oblique cultural transmission ($\gamma > s$). A high level of assortative mating of
A-types would work as well, but we do not explore that possibility here.

11.6 Why is Internalization Socially Beneficial?

Internalized norms need not enhance the fitness of group members. Indeed,
many social norms are costly, such as those involving invidious displays
of physical prowess and promoting costly wars and feuds (Edgerton 1992).
The reason for the feasibility of antisocial norms is that once the internaliza-
tion allele has evolved to fixation, there is nothing to prevent group-harmful
phenotypic norms from also emerging, provided they are not excessively
costly to the individual, given the strength of the payoff-based updating
process ($\eta$). The evolution of these harmful norms directly reduces the
overall fitness of the population.

Yet, as Brown (1991) and others have shown, there is a tendency in vir-
tually all successful societies for cultural institutions to promote social and
eschew anti-social norms, and for A-types to embrace these social norms.
The most reasonable explanation for the predominance of socially benefi-
cial norms is weak group selection: societies that promote social norms
have higher survival and reproduction rates than societies that do not.

Weak group selection (§6.1) is sufficient for the proliferation of socially
beneficial norms as long as the conditions for the stability of the altruistic
equilibrium (11.7) are met. A-types in groups at or near this equilibrium will be as fit as other members of their groups and will therefore not suffer adverse within-group selection. But, the fitness of all members of groups at or near the altruistic equilibrium will exceed that of members of groups that support group-harmful norms. The evolutionary dynamic is thus an equilibrium selection problem with differential group survival favoring the selection of the altruistic equilibrium.

The question of interest, then, is whether the updating system captured by our vertical, oblique and payoff based transmission is itself likely to evolve such that the condition for the stability of the altruistic equilibrium (11.7) will be satisfied. If groups with strong systems of oblique transmission (i.e., high levels of $\gamma$) were to do poorly for some reason, then 11.7 might not be satisfied in a long-term evolutionary dynamic. Recall that the in Chapters 8 and 9 we asked a similar question. Having shown that culturally transmitted reproductive leveling and within-group segmentation practices favor the evolution of a genetically transmitted altruistic predisposition (Chapter 8) and that intergroup hostilities are essential to this process (Chapter 9) we asked if these altruism-favoring conditions themselves could evolve. Here, instead we explore the coevolution of three distinct aspects of a population: the dynamics of the distributions of its genotypes and phenotypes and the evolution of the process by which individuals update their socially learned traits. The third will require an exploration of the dynamics of $\gamma$, the effectiveness of its institutions of socialization and $\eta$, the effect of payoff differences in inducing individuals to switch from altruist to selfish types. As we did in Chapters 8, 9, and 10, we will also determine if an initially rare altruistic trait can proliferate in a reasonable time frame, and if it is sustained in a stochastic environment.

Given the complexity of this task, selection on genes, learned behavior, and two aspects of a society’s social learning system operating at both the individual and group level, we are not able to develop an illuminating analytical model, and so, as in previous chapters we created an agent-based model of society with the following characteristics (the specific assumptions made are not critical, unless otherwise noted). The society consists of 1000 groups, each initially comprising 12 members per generation, or a census size of about 36, the typical size of a Pleistocene hunter-gatherer group, arranged spatially on a torus (a 50 x 50 inner-tube-type grid with the opposite edges identified). Each group started out with 2% $aAC$ types, 1% $aAD$ types, 1% $aSC$ types, 1% $aSD$ types, and 95% $bSD$ types. Ta-
Simulation Parameter | Value
---|---
Initial Frequency of $a_{AC}$ | 2%  
Initial Frequency of $a_{AD,aSC,aSD}$ | 1%  
Initial Frequency of $b_{SD}$ | 95%  
Fitness cost of altruism $s$ | 0.03  
Gain from internalizing fitness-enhancing norms $f$ | 0.06  
Fitness cost of internalization physiology $u$ | 0.01  
Initial Range of Rate of Oblique Transmission $\gamma$ | [0,0.9]  
Initial Range of Imitation Rate $\eta$ | [0,0.9]  
Initial Group Size | 12  
Conflict Rate | 10%  
Cost of $\gamma$ | 5$s$  
Cost of $\eta$ | 5$s$  
Fitness contribution of A-type to group | 0.05  
Mutation Rate | 0.01%  
Migration Rate | 25%  
Number of Groups | 1000

Table 11.6. Parameters for the Simulation of the Spread of Strong Reciprocity Through Weak Group Selection. \([a, b]\) signifies the initial seeding of the groups with values drawn from the uniform distribution on \([a, b]\). The values of $s$, $f$, $u$, as well as the fitness contribution of A-types, the mutation and migration rates are the same and unchanging for all groups and all generations.

Table 11.6 summarizes the parameter choices of the simulation. In all groups, $s = 0.03$, $f = 0.06$ and $u = 0.01$, because they represent individual-level costs and benefits unrelated to any group differences in social structure. We take $s$ as constant because we are not concerned with the obvious point that groups with higher $s$ will be disadvantaged. We also fixed the benefit of altruism $a 0.05$ for all groups; i.e., a group of all A-types has a 5% fitness advantage over a group of all non-altruists.

By contrast, the extent $\gamma$ of oblique transmission is clearly a socially-determined variable, societies with higher $\gamma$ according more social influence to A-type elders. As in Chapter 8, we assume that institutions are not free goods. In this case a more effective socialization system (greater $\gamma$) comes at the price of a larger fitness disadvantage for the A-types. The time they spend teaching altruistic behavior, for example, they cannot be seeking out mating opportunities and caring for their offspring.
We set the cost of per A-type of $\gamma$ to be $s(\gamma)$; i.e. setting $\gamma = 0.80$ in a group is equivalent to raising the fitness cost to A-types by $0.8s$. We found that this cost is inversely related to the long-run value of $\gamma$, as one might expect. The level of $\eta$, the lure of higher payoffs in motivating the regression from altruism to self-interest, is also socially determined. A-types, whose numbers are reduced by desertion to self-interest when $\eta$ is substantial, can devote time and energy to reducing the lure of payoffs, teaching, for example the value of non-material well-being. To reflect his we imposed a cost of $s(1 - \eta)$ on the A types. Thus setting $\eta = 0.20$ in a group is equivalent to raising the fitness cost to A-types by $0.8s$.

Each group initially was randomly assigned a value of $\gamma$ and a value of $\eta$. Random variation in social learning arrangements ("institutional mutation") allowed $\eta$, and $\gamma$ to increase or decrease by 1% of their values. The migration rate was set to 25% per generation (very high for a genetic model but reasonable for a cultural model), and the mutation rate was set to 0.01% per generation, and migration was always to a neighboring group, individuals taking their phenotypic traits with them.

In each generation, for each of the 1000 groups, we simulated the model as described in the previous sections, and update the frequencies of the various types in each group, according to the fitness effect of their A phenotype and the fraction of the group that exhibits this phenotype. Then, 25% of individuals in each group migrated randomly to neighboring groups, bringing their phenogenotype with them. Selection among groups takes two forms in this model. First, if group size drops below a minimum (set to one third of initial group size, or four), it is replaced by a copy of the neighboring group that has the highest average fitness of group members. Second, with a small probability each generation, a group enters into conflict with another randomly chosen group. The group with higher fitness prevails, and members of the losing group copy the parameters of members of the winning group.

We ran this model many times with varying numbers of generations, and varying the parameters described above. The system always stabilized rapidly, there is virtually no variation in final values across runs, the specific assumptions concerning the parameters move in the intuitively expected direction, and were never critical. The parameter values always allow zero altruism to be a stable evolutionary equilibrium, but with as few as 2% initial A-types, altruism always stabilized at a high level. A run with the parameters described above is exhibited in Figure 11.1. There is always strong selection on the rate of oblique transmission, unless the cost of maintaining
Figure 11.1. The Evolution of Endogenous Parameters. The Rate of Oblique Transmission refers to $\gamma$, and the rate of imitation to $\eta$ in the model.

$\gamma$ at a high level is extremely high (about 10s). Selection for lower $\eta$ is also quite strong, so a high cost of reducing it is needed to prevent $\eta$ from falling to very low levels in the long run.

Figure 11.1 shows the evolution of the endogenous parameters in this simulation. The fraction of A-types increases to about 57% by the end of the run. This value varies between 50% and 75%, depending on the costs, borne by A-types alone, for maintaining a high $\gamma$ and a low $\eta$. It is clear that all three parameters of the model undergo strong selection, $\gamma$ rising to 0.083, and $\eta$ falling to 0.26 ($\gamma$ is multiplied by 10 in the figure).

Migration does not undermine the altruistic equilibrium, because most of the effects occur on the cultural rather than the genetic level, and migrants respond to the social learning environment of their new home.

The simulation thus identifies a wide range of parameter values under which a system of cultural transmission biased towards socialization of the young for altruism and minimizing the lure of material payoffs could itself evolve, and if it did that these social learning arrangements would support a frequency of altruism in the population.
11.7 The Programmable Brain

Vertical, oblique, and payoff-based updating all involve the internalization of norms, which is itself a costly process \((u > 0)\). Taking on a general rule of behavior as an objective rather than a constraint or an instrument towards some other end is likely to be costly, as the rule will not be ideally suited to all situations, and its internalization deprives the individual of flexibility in dealing with such situations on a case-by-case basis. The parochial preferences that motivate the exclusion of outsiders studied in Chapter 9 is an example of a personally costly general rule of behavior. Moreover, a considerable fraction of the total available time of the members of most societies is spent teaching the young the proper way to behave, rather than providing for the nutritional and other needs of its members.

Why, then, are humans so susceptible to internalizing general rules? If this susceptibility were subject to a purely payoff-based selection process, whether fitness- or payoff-sensitive, one might expect it to be eliminated from any population in which it appeared. Why then are general rules of behavior common? An answer that we have found persuasive (Heiner 1985) is that internalizing general rules of behavior may persist in an evolutionary dynamic because it relieves the individual from calculating the costs and benefits in each situation and reduces the likelihood of making costly errors. A similar argument led John Stuart Mill to remark, “Being rational creatures [sailors] go to sea with it [the Nautical Almanac] already calculated; and all rational creatures go out upon the sea of life with their minds made up on the common questions of right and wrong, as well as on many of the far more difficult questions of wise and foolish” Utilitarianism(1861):31.

Our models show that cultural transmission and the capacity to internalize norms may coevolve if some of these norms are fitness-enhancing for the individuals who adopt them. But, if this is the case, what is the evolutionary advantage of taking on the costs of socialization and internalization?

Like other animals, our body produces the sensations of pleasure and pain in response to the things we experience, and this is what induces our behavior. These hedonic responses that constitute the proximate causes of behavior can be represented as what we have in Chapter 5 defined as preferences: reasons for behavior, other than beliefs and capacities, that account for the actions an individual takes in a given situation. These preferences are subject to natural selection, as well as social learning in some animals, and there is some reason to think that, for most animals most of the time,
preferences induce behavior approximating that would result if the individual animal were to deliberately maximize its fitness, at least locally.

Cultural transmission and internalization make humans an exception to this general proposition. Cultural transmission and internalization affect our hedonic responses to situations and induce behaviors that may diverge substantially and systematically from what an individual fitness maximizer would do. As we saw in the Introduction to this chapter, individual and even average fitness-reducing behaviors can be successfully promoted by cultural transmission and internalization. But the internalization of culturally transmitted norms can also do better than natural selection in inducing behaviors that enhance fitness. This is true for two reasons.

First, except under special circumstances, individual fitness maximization does not maximize average fitness of the members of a group. The elimination of A-types from a random mixing population as a result of a fitness-based payoff-based updating dynamic is a pertinent example. Other examples were studied in chapters 8 and 10. This being the case, groups that override individual fitness-maximizing by means of the cultural transmission of internalized norms may experience higher group average fitness than other groups. These group benefits may offset the costs just mentioned. Indeed, this is one of the key dynamics accounting for the emergence of altruism in the above models, and of social preferences in general.

In our model of socialization, oblique transmission converts a fraction of self-regarding types into altruists. But we did not ask about the proximate motives for the altruists helping others. Does oblique transmission work by teaching children the golden rule or Kant’s categorical imperative? By warning them that God may be watching?

These and other cognitive reasons for good behavior are no doubt involved, but the motivation to help others and to act ethically often short-circuits these reflective processes in favor of more visceral influences on behavior such as anger, shame, elation and guilt. To readers who share our horror of road rage and honor killings, the claim that visceral reactions are among the proximate motives for generous, fair-minded and civic actions may seem surprising. But it is true, and we think that a good case can be made that the social emotions evolved precisely because they motivated prosocial actions.
Let’s not forget that the little emotions are the great captains of our lives and we obey them without realizing it.

Vincent Van Gogh, *Letter to his brother Theo* (1889)

The heart has reasons that Reason knows nothing about.

Blaise Pascal, *Pensées* (1995[1670])

Social emotions, love, guilt, shame, self-righteousness, and others, are responsible for the host of civil and caring acts that enrich our daily lives and render living, working, shopping, traveling among strangers, even conducting scientific research, feasible and pleasant. Adherence to social norms is underwritten not only by cognitively mediated decisions, but also by emotions (Frank, 1987, 1988; Ekman, 1992; Damasio, 1994; Elster, 1998; Boehm 2007). When Bosman and Zeelenberg (2001) assayed the feelings of respondents in an ultimatum game, they found that low offers provoked anger, contempt and sadness, that the intensity of the self-reported emotions predicted the respondents' behavior, stronger emotions inducing rejections of low offers. Interestingly, the introduction of an hour-long “cooling off” period between offer and the respondent’s choice of an action had no effect on either reported emotions or on the rejection behaviors of the respondents. Recall from Chapter 5 that Sanfey et al. (2003) found that those rejecting low offers in an ultimatum game experienced heightened levels of activation in the brain areas associated with disgust and anger.

One of the most important emotions sustaining cooperation is shame, the feeling of discomfort at having done something wrong not only by one’s own norms but also in the eyes of those whose opinions matter to you. Shame differs from guilt in that, while both involve the violation of a norm, the former but not the latter is necessarily induced by others’ knowing about the violation and making their displeasure known to the violator.

We will suggest that shame, guilt, and other social emotions may function like pain, in providing personally beneficial guides for action that by-
pass the explicit cognitive optimizing process that lies at the core of the
standard behavioral model in economics and decision theory. Pain is one
of the six so-called ‘basic’ emotions, the others being pleasure, anger, fear,
surprise, and disgust. Shame is one of the seven so-called social emotions,
of which the others are love, guilt, embarrassment, pride, envy, and jealousy
(Plutchik 1980, Ekman 1992). Basic and social emotions are expressed in
all human societies, although their expression is affected by cultural condi-
tions. For instance, in all societies one may be angered by an immoral act,
or disgusted by an unusual foodstuff, but what counts as an immoral act or
a disgusting foodstuff is, at least to some extent, culturally specific.

Antonio Damasio (1994):173 calls an emotion a “somatic marker,” that
is, a bodily response that “forces attention on the negative outcome to which
a given action may lead and functions as an automated alarm signal which
says: Beware of danger ahead if you choose the option that leads to this
outcome... the automated signal protects you against future losses.” Emo-
tions thus may contribute to the decision-making process by working with,
not against, reason. Damasio continues, analogizing emotions to physical
pain: “suffering puts us on notice... it increases the probability that indi-
viduals will heed pain signals and act to avert their source or correct their
consequences.” (p. 264)

To explore the role of guilt and shame in inducing social behaviors we
will consider a particular interaction having the structure of a public goods
game. In the public good setting, contributing too little to the public ac-
count may evoke shame if one feels that one has appropriated “too much”
to oneself. Because shame is socially induced, being punished when one
has contributed little triggers the feeling of having taken too much. In this
case, the effect of punishment on behavior may not operate by changing
the incentives facing the individual, that is by making it clear that his pay-
offs will be reduced by the expected punishments in future rounds. Rather it
evokes a different evaluation by the individual of the act of taking too much,
namely, shame. This is the view expressed by Jon Elster (1998):67 “ma-
terial sanctions themselves are best understood as vehicles of the emotion
of contempt, which is the direct trigger of shame.” Thus, self-interested
actions, per se, may induce guilt, but not shame. If one contributes little
and is not punished, one comes to consider these actions as unshameful.
If, by contrast, one is punished when one has contributed generously, the
emotional reaction may be spite towards the members of one’s group.
We assume individuals maximize a utility function that captures five distinct motives: one’s individual material payoffs, how much one values the payoffs to others, this depending both on ones’ altruism and one’s degree of reciprocity, and one’s sense of guilt or shame in response to one’s own and others’ actions. To this end, we will amend and extend a utility function derived from the work of Geanakoplos et al. (1989), Levine (1998), Sethi and Somanathan (2001), and Falk and Fischbacher (2006).

In Chapter 5, we presented experimental evidence consistent with the view that punishment not only reduces material payoffs of those who transgress norms, but also may recruit emotions of shame towards the modification of behavior. Indeed, we showed in §5.6 that in some societies many defectors react to being punished by increasing their contribution to the group, even when the punishment does not affect material payoffs, consistent with the shame response, while in other societies they react by counter-punishing contributors, consistent with an anger response. Social emotions in response to sanctions can thus either foster or undermine cooperation. Reacting to sanctions, then, is often not a dispassionate calculation of material costs and benefits, but rather involves the deployment of culturally specific social emotions. In Chapter 10 we showed that the altruistic punishment of shirkers by strong reciprocators can proliferate in a population and sustain high levels of cooperation, but we tacitly assumed that those punished would react prosocially rather than antisocially. Here, we focus on the manner in which social emotions and punishment of miscreants may be synergistic, each enhancing the effects of the other.

In §12.1, we model the process by which an emotion such as shame may affect behavior in a simple public goods game. We then show that shame and guilt along with internalized ethical norms allow high levels of cooperation to be sustained with minimal levels of costly punishment, resulting in mutually beneficial interactions at limited cost. In §12.2, we ask how prosocial emotions such as shame might have evolved. Indeed, we show that if the return to the public good is sufficiently high and if the reciprocity motive is strong, an individual is benefitted by increasing his shame parameter, as this leads to an outcome in which his net utility is higher.
12.1 Reciprocity, Shame, and Punishment

Consider two individuals who play a one-shot public goods game in which each has a norm concerning the appropriate amount to contribute to the public project, and each (a) values his own material payoff, (b) may prefer to punish others who contribute insufficiently, (c) feels guilt if he contributes less than the norm; and finally (d) experiences shame if he is sanctioned for having contributed less than the norm. This psychological repertoire captures some of the motives that explain cooperation in behavioral experiments. The results that follow generalize to an $n$-person interaction.

We assume each individual starts with a personal account equal to 1 unit. Each individual contributes to the public project and amount $a_i$, $0 \leq a_i \leq 1$, and each receives $\chi (a_1 + a_2)$ from the project, where $1/2 < \chi < 1$. Thus, the individuals do best when both cooperate $(a_i, a_j = 1)$, but each has an incentive to defect $(a_i, a_j = 0)$ no matter what the other does. In the absence of punishment, this two-person public goods game thus would be a prisoner’s dilemma. But at the end of this cooperation period there is a punishment period, in which the individuals are informed of the contribution of the other individual, and each individual may impose a penalty $\mu$ on the other individual at a cost $c(\mu) = c\mu^2/2$. This, and the other functional forms below, are chosen for expositional convenience.

In what follows, we represent the two players as $i$ and $j$, where $j \neq i$. Letting $\mu_{ij}$ be the level of punishment of individual $j$ by individual $i$, the material payoffs to $i$ is then given by

$$
\pi_i = 1 - a_i + \chi (a_1 + a_2) - \mu_{ji} - c(\mu_{ij})
$$

(12.1)

In 12.1, the first two terms give the amount remaining in $i$’s private account after contributing, the third term is $i$’s reward from the public project, the fourth term is the punishment inflicted by $j$ upon $i$, and the final term is the cost to $i$ of punishing $j$.

We assume that the norm is that each should contribute the entire endowment to the public project. The results generalize to the case where the norm is less stringent. Individual $i$ may wish to punish $j$ by reducing $j$’s payoffs, if $i$ is a reciprocator and $j$ contributes less than the entire endowment. To represent the propensity of $i$ to punish $j$ for not contributing, we assume that $i$’s valuation of $j$’s payoff is

$$
\beta_{ij} = \lambda_i (a_j - 1),
$$

(12.2)
where we assume $0 < \lambda_i < 1$, so that unless $j$ contributed his entire endowment, $i$ receives a psychic benefit from lowering $j$’s material payoff that is proportional to $j$’s shortfall. The parameter $\lambda_i$, $0 < \lambda_i < 1$, is the strength of $i$’s reciprocity motive. The condition that $\lambda_i < 1$ ensures that individual $i$ cannot value $j$’s payoffs negatively more than he values his own positively thus should both payoffs increase proportionally, individual $i$ cannot be worse off.

The shame experienced by $i$ is a psychic cost proportional to the product of the degree to which he is punished by $j$, and the extent to which his contribution falls short of the norm, and is equal to $\gamma_i (1 - a_i) \mu_{ij}$. Thus, punishment triggers shame, which is greater the more the individual has kept for himself rather than contributing to the public project, and the larger is $\gamma_i$, the susceptibility of individual $i$ to feeling shame. Finally, $i$ may feel guilt simply for having violated his internal standards of moral behavior. We represent this feeling by $-\gamma_i (1 - a_i)$, which is negative for $\gamma_i > 0$ unless $i$ contributes the full amount to the project.

The utility function $i$ is then given by

$$u_i = \pi_i + \beta_{ij}(1 - a_j + x(a_1 + a_2) - \mu_{ij}) - (\gamma_i + \nu_i \mu_{ij})(1 - a_i)$$

(12.3)

We have not included the cost to $j$’s of punishing $i$, in the material payoffs of $j$ that $i$ takes account of when choosing his contribution level. Note that shame and punishment are complementary in the sense that an increase in the susceptibility of shame increases the marginal effect of punishment on the individual’s utility, and an increase in the level of punishment similarly raises the marginal effect of a shameful action on the actor’s utility. Shame thus enhances what is termed the ‘punishment technology,’ the effectiveness of which is measured by the ratio of the utility loss inflicted on the target, including both the subjective costs and the reduction in payoffs from equation (12.1), to the marginal cost to the punisher of undertaking the punishment. This punishment effectiveness ratio for $i$’s punishment of $j$ is thus

$$1 + \nu_j (1 - a_j)$$

(12.4)

$\frac{c \mu_{ij}}{c \mu_{ij}}$

from which it is clear that the punishment of $j$ is more effective the more susceptible to shame is $j$.

Because each individual’s valuation of the payoffs of the other depends on the actions the other takes, it is clear that the actions taken by each will be mutually determined. For any given value of $j$’s action, there will be
an action—a best response—by \( i \) that maximizes his utility as expressed in equation 12.3. The best response function for individual \( i \) is shown in Figure 12.1, along with the analogous best response function for \( j \). Their intersection is the mutual best response, and is therefore the Nash equilibrium. In Figure 12.1 we see that the best response \( a_i \) is an increasing function of \( a_j \), and the \( a_i \) schedule shifts up when \( v_i, \gamma_i \) or \( \lambda_j \) increases, corresponding to our intuitions concerning the model. There is also a minimal level of susceptibility to shame supporting positive contributions. The minimal level of shame that will induce a positive contribution is increasing in the cost of punishment, and decreasing in \( i \)’s susceptibility to guilt \( \gamma_i \), \( j \)’s level of reciprocity \( \lambda_j \), and the productivity \( \chi \) of the public project, again confirming our intuitions.

\[
\begin{align*}
\text{Figure 12.1.} & \quad \text{Mutual determination of contributions to a public project.} \\
& \quad \text{Note:} \\
& \quad \text{The functions slope upwards because the individuals are reciprocators and shift as shown when susceptibility to shame, } v \text{, increases, because this enhances the effects of punishment. There is no reason to think that the function would take the linear form shown here.}
\end{align*}
\]

Suppose the level of shame of both individuals were to increase. This is shown in Figure 12.1 by the dashed lines. The result is a displacement of the mutual best response so that both individuals contribute more, and as a result the level of punishment is less. This is the sense in which we mean that because shame enhances the effectiveness of punishment: it economizes on the cost of punishment. When one individual’s susceptibility to shame increases the other individual benefits and when this occurs for both, as in Figure 12.1, both benefit. Payoffs therefore are higher in a population that has inculcated a sense of shame in its members, as could be the case.
for example through the kinds of population-wide internalization of norms studied in the previous chapter.

But could an enhanced sense of shame raise an individual’s payoffs even if the other individual’s sense of shame remained unchanged? Having studied the evolution of altruistic punishment in Chapter 10 and wanting to focus here on shame, we have assumed that the direct effect of increasing $i$’s own contribution lowers $i$’s payoffs, even taking account of the reduced punishment that $i$ will receive from $j$ as a result; namely, $1 - \chi > \lambda /c$. Thus the only way that $i$’s payoffs can rise as a result of an increase in his susceptibility to shame if $j$’s reciprocal response to $i$’s increase contribution is sufficiently great. We can indeed show (§A12) that if the common project is sufficiently productive ($\chi$ is large) and if individuals are sufficiently reciprocal ($\lambda$ is large), an enhanced level of shame may raise one’s own payoffs. This cannot occur, however, where reciprocity is absent or where the benefits to cooperation are minimal, or where shame itself at a very low level. It is possible, then, that reciprocity and shame might have coevolved.

### 12.2 The Evolution of Social Emotions

Human behaviors systematically deviate from the model of the selfinterested actor, and we think the evidence is strong that social emotions account for much of the discrepancy. But, this description of behavior would be more compelling if we understood how social emotions might have evolved, culturally, genetically, or both. There are two puzzles here. First, social emotions are often altruistic, indicating actions benefiting others at a cost to oneself, so that any dynamics in which the higher payoff trait tends to increase in frequency, social emotions would eventually disappear. We have addressed this puzzle in the previous four chapters, showing that by the process of group competition, reproductive leveling, and norm internalization, vertically transmitted altruistic traits may evolve.

The second puzzle concerns social emotions *per se*. How could it ever be evolutionarily advantageous to bypass one’s cognitive decision making capacities and let behavior be influenced by the visceral reactions associated with one’s emotions? We have addressed a similar question in the previous chapter: internalizing norms may be a way of economizing the costs of calculating benefits and costs in each situation, and of averting costly er-
rors when the calculations go wrong. A related argument, we think, helps explain the evolutionary viability of social emotions.

Humans tend to be present-oriented, a condition they share with other animals (Stephens et al. 2002). We tend to discount future costs and benefits myopically, that is, more than either a fitness-based or a lifetime welfare-based accounting would require. The mismatch between our impatience and our fitness is in part due to the payoff to patient behaviors that resulted from the extended life histories and prolonged period of learning the skills associated with the distinctive skill-intensive human feeding niche based on hunted and extracted foods. Prior to this period in human history, the importance of the future was more limited and largely concerned the survival of one’s offspring. A genetically transmitted disposition to assist one’s relatives may have produced a selective degree of patience as a byproduct of inclusive fitness maximization, resisting stealing food from one’s offspring, for example. Even if our genetic development in a cooperative social context has mitigated the extreme short-term benefits of lying, cheating, killing, stealing, and satisfying immediate bodily needs, such as wrath, lust, greed, gluttony, sloth, we nevertheless have a fitness-reducing bias towards behaviors that produce immediate satisfaction at the expense of our long-run well-being.

The internalization of norms and the expression of these norms in a social emotion such as guilt addresses this problem by inducing the individual to place a contemporaneous value on the future consequences of present behavior, rather than relying upon an accurate accounting of its probable payoffs in the distant future. One curbs one’s anger today not because there may be harmful effects next month, but because one would feel guilty now if one violates the norms of respect for others and the dispassionate adjudication of differences. One punishes others for behaving anti-socially not because there are future benefits to be gained thereby, but because one is angered at the moment.

Do the social emotions thus function in a manner similar to pain? Complex organisms have the ability to learn to avoid damage. A measure of damage is pain, a highly aversive sensation the organism will attempt to avoid in the future. Yet an organism with complete information, an unlimited capacity to process information, and with an fitness-maximizing way of discounting future costs and benefits would have no use for pain. Such an individual would be able to assess the costs of any damage to itself, would calculate an optimal response to such damage, and would prepare optimally
for future occurrences of this damage. The aversive stimulus, pain, could then be strongly distorting of optimal behavior. If you sprain your ankle while fleeing from a lethal predator, you might have a better chance of survival if you could override the pain temporarily. Because pain per se clearly does have adaptive value, it follows that modeling pain presupposes that the individual experiencing pain must have incomplete information and/or a limited capacity to process information, and/or an excessively high rate of discounting future benefits and costs. Are guilt and shame social analogues to pain?

If being socially devalued has fitness costs, and if the amount of guilt or shame that a given action induces is closely correlated with the level of these fitness costs that would otherwise not be taken account of, then the answer is affirmative. The same argument will hold not only for fitness costs, but for any effect, possibly operating through cultural transmission, that reduces the number of replicas an individual will generate. Shame and guilt, like pain, are aversive stimuli that lead the individual experiencing them to repair the situation that led to the stimulus, and to avoid such situations in the future.

### 12.3 The Great Captains of our Lives

Shame and guilt, like pain, dispense with an involved optimization process by means of a simple message: whatever you did, undo it if possible, and do not do it again. Two types of selective advantage thus may account for the evolutionary success of shame and related social emotions. First, social emotions may increase the number of replicas, by either genetic or cultural transmission, of an individual who has incomplete information (e.g., as to how damaging a particular anti-social action is), limited or imperfect information-processing capacity, and/or a tendency to undervalue costs and benefit that accrue in the future. Probably all three conditions conspire to induce people to respond insufficiently to social disapproval in the absence of social emotions. The visceral reactions associated with these emotions motivate a more adequate response, one that will avert damage to the individual. Of course the role of social emotions in alerting us to negative consequences in the future presupposes that society is organized to impose those costs on norm violators. The social emotions may thus may have coevolved with the reciprocity-based emotions motivating punishment of antisocial actions, modeled in the previous chapters.
The second selective advantage favoring the evolution of social emotions refers specifically to shame. The fact that higher levels of shame among members of a group, the higher (in equilibrium) will be the sum of their payoffs also suggests that shame may evolve through the effects of group competition. As we have seen, where the emotion of shame is common, punishment of antisocial actions will be particularly effective and as a result seldom used. Thus groups in which shame is common can sustain high levels of group cooperation at limited cost and will be more likely to survive environmental, military and other challenges, and thus to populate new regions or sites vacated by groups that failed.

As a result, selective pressures at the group level will also favor religious practices and systems of socialization that support susceptibility to shame for failure to contribute to projects of mutual benefit of the type modeled in the previous two sections. Moreover, where the returns to cooperation and levels of reciprocity are sufficiently great, an individual who acquires an enhanced sense of shame may increase his own individual payoff. Under these conditions an individual that acquired enhanced shame by chance (a mutation, developmental accident or other) could invade a large population of individuals with lesser levels of shame. Thus, a genetic or cultural predisposition to shame could increase in a population even in the absence of group competition.

It is quite likely then, that the ‘moralistic aggression’ that is involved in the altruistic punishment of miscreants and that motivated the punishment of shirkers in Chapter 10 also created a selective niche favorable to the emergence of shame and other social emotions, or what Christopher Boehm calls a conscience:

The human conscience evolved in the Middle to Late Pleistocene as a result of subsistence turning to the hunting of large game. This required...cooperative band-level sharing of meat...bands had to gang up physically against their alphas to ensure efficient meat distribution. This set the stage for morality to develop as a new, more socially-sensitive type of personal self-control became adaptive for individuals living in these punitive groups. Thus a conscience began to develop biologically. In turn...conscience transformed social control by making punitive sanctioning increasingly moral and also less lethal, as group ostracism and shaming evolved.
Combining the model of this chapter and that of Chapter 10, the emergence of shame would have reduced the costs of punishing transgressors incurred by the strong reciprocators. The reason for this is that gossip and ridicule could then suffice where physical, often violent, elimination from the group had been necessary in the absence of shame. The proliferation of strong reciprocators engaging in altruistic punishment that this cost reduction allowed would then have enhanced the advantages of shame.

Thus the moralistic aggression motivating the altruistic punishment of defectors, and shame, may have coevolved, each providing the conditions favoring the proliferation of the other. The groups in which this occurred initially, perhaps among our foraging ancestors in Africa, would have enjoyed survival advantages over other groups.
13

Conclusion: Human Cooperation and its Evolution

This is the gist of human psychology... what the hero does all feel that they ought to have done as well. The sophisms of the brain cannot resist the mutual aid feeling, because this feeling has been nurtured by thousands of years of human social life and hundreds of thousands of years of pre-human life in societies.


Any animal whatever, endowed with well-marked social instincts, the parental and filial affections being here included, would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well developed, or nearly as well developed, as in man.


About 55,000 years ago, a group of hunter-gatherers left Africa and began to move eastward along the shores of the Indian Ocean. They may have originated in the Upper Rift Valley in modern day Kenya. They could have been the descendants of the cooperative early humans we described at the outset, living 30,000 years earlier at the mouth of the Klassies River far to the south. Wherever they came from, they eventually crossed hundreds of kilometers of open ocean before reaching Australia, just 15,000 years later. We do not know if they encountered or simply bypassed communities of Homo floresiensis who persisted in what is now Indonesia almost to the end of the Pleistocene. Others crossed the Levant and somewhat later occupied Europe, quickly driving the Neanderthals to extinction. Though the possibility of multiple human origins cannot be eliminated, it is now widely thought that the descendants of this small group eventually peopled the en-
tire world and are the ancestors of all living humans (Klein 1999, Foley 1996).

This second great exodus from Africa is remarkable for its speed and eventual spread. One cannot resist speculating about the capacities that made these particular individuals such lethal competitors for the (also large-brained and tool-making) Neanderthals or that allowed the construction of ocean-going craft. Some attractive candidates can be ruled out. The physiological innovations allowing for more effective speech, rearrangement of breathing and digestive tracts for example, had occurred much earlier. Likewise, the dramatic expansion of hominid brain size had occurred before two million years ago. Richard Klein suggests a “selectively advantageous mutation” that facilitated the cultural transmission of behaviors as a possible cause.

Arguably this was the most significant mutation in the human evolutionary series for it produced an organism that could radically alter its behavior without any change in its anatomy and that could cumulate and transmit alterations at a speed that anatomical innovation could never match (Klein 2000):18.

But, as Klein himself points out, the only evidence for such a super-mutation are the facts it is intended to explain (Klein 2000). Whether the source was a single revolutionary innovation or, as many now think (McBrearty and Brooks 2000) the result of a long process of incremental changes, the linguistic capacities and the cultural transmission of norms of social conduct that supported cooperation were a necessary part of the human repertoire that made the peopling of the world possible. These same capabilities must be part of any account of the remarkable success of humans as a species then and since.

13.1 The Origins of Human Cooperation

Humans became a cooperative species because our distinctive livelihoods made cooperation within a group highly beneficial to its members and, exceptionally among animals, we developed the cognitive, linguistic and other capacities to structure our social interactions in ways that allowed altruistic cooperators to proliferate.

Human reliance on the meat of large hunted animals and other high quality, large package size, and hence high variance, foods meant that our
livelihoods were risky, skill-intensive, and characterized by increasing returns to scale. Deploying skills that require years to acquire favored the evolution of large brains, patience, and long lives (Kaplan et al. 2000, Kaplan and Robson 2003). Organizing and sharing the returns to successful hunting additionally favored groups that developed practices of sharing information, food, and other valued resources (Boehm 2000). Moreover, the long period of dependency of human offspring on adults, in part the result of the prolonged learning curve associated with hunting and gathering, meant that there were substantial benefits to cooperative child-rearing practices extending beyond the immediate family. Prolonged juvenile dependency also generates a net food deficit for families with adolescent children, greatly increasing the benefits of food sharing among unrelated individuals and other forms of social insurance (Kaplan and Gurven 2005). Our experimental evidence, presented in Chapter 5, shows that among today’s small scale societies, those that are especially reliant on big game, like the Lamalera whale hunters that we studied in Indonesia, and those for whom livelihoods require either joint efforts in acquisition or sharing in distribution, are especially likely to exhibit the social preferences that underpin altruistic cooperation.

One of the reasons for the connection between the potential benefits of cooperation and the prevalence of cooperative behaviors that we discovered in our models and simulations is that where the benefits associated with cooperation relative to the costs are substantial, it is more likely that the evolutionary processes of gene-culture coevolution will support populations with large numbers of cooperators, whether altruistic or mutualistic. A high ratio of benefits to costs makes cooperation an evolutionary likely outcome (to use Robert Boyd’s phrase) because, as our models and simulations, for example Figures 3.3, 3.4, 10.1, and 10.4 confirmed, in virtually any plausible evolutionary dynamic in which stochastic shocks to payoffs and to behaviors play an important role, the likelihood that a population will develop and maintain cooperative practices is higher, the greater are the net benefits of cooperation.

But the fact that cooperation was group-beneficial in the environments of early humans does not explain why it evolved, for individuals bear the costs of their cooperative behaviors, while it is often others who enjoy the benefits. Thus, the distinctive human livelihood and associated cognitive capacities and longevity are necessary but not sufficient to explain the extent and nature of human cooperation. While benefits of cooperation accru-
ing to the individual cooperator may sometimes offset the costs, this is not likely to have been the case in many situations in which cooperation was essential to our ancestors, including defense, predation and surmounting environmental crises. In these situations involving large numbers of individuals facing their possible demise, people with self-regarding preferences would not cooperate, irrespective of their beliefs about what others would do. As a result, for cooperation to be sustained, social preferences would have to motivate at least some of those involved.

The distinctive human capacity for institution-building and cultural transmission of learned behavior allowed social preferences to proliferate. Our ancestors used their capacities to learn from one another and to transmit information to create distinctive social environments. The resulting institutional and cultural niches reduced the costs borne by altruistic cooperators and raised the costs of free-riding. Among these socially constructed environments, three were particularly important: group-structured populations with frequent and lethal inter-group competition, within-group leveling practices such as sharing food and information, and developmental institutions that internalized socially beneficial preferences.

These culturally transmitted institutional environments created a social and biological niche favorable to the evolution of the social preferences on which altruistic cooperation is based. We can only speculate, of course, about the initial appearance and proliferation of these preferences. But, there are two reasons why their emergence was highly likely. The first is that the preferences that constitute strong reciprocity and some other social preferences could appear de novo as the result of only a small behavioral modification of either kin-based altruism or reciprocal altruism, behaviors that support cooperation on the basis of self-regarding preferences. In the case of kin altruism, those behaving altruistically toward kin may have simply ceased discriminating against the non-kin members of their groups. Likewise, a reciprocal altruist could become a strong reciprocator by simply fostering the compliance with norms not only with the family but in the larger society.

The second reason why the emergence of social preferences among early humans would be highly likely is the small size and vast number of foraging bands during the Late Pleistocene and earlier. Even if strong reciprocity initially emerged in very small fraction of the human population, it is highly likely that over tens of thousands of generations and more than a quarter of a million of foraging bands, it would have occurred that the
strong reciprocators or other altruistic cooperators were prevalent in one or more such groups at some point. These bands would have done very well in competition with other bands.

We have sought to explain how humans came to develop these exceptional social preferences and the cooperative social practices that supported them, taking the distinctive nature of human ecology, diet, and life course as pre-existing. This analytical simplification is almost surely historically inaccurate. The distinctive nature of human livelihoods, the importance of hunted and extracted as opposed to collected foods, apparently does not predate and is not the cause of the emergence of cooperation. Rather, it appears that the two developed in tandem.

Though we have not addressed this question, we think it likely that the models presented here, suitably amended, would illuminate the coevolution of human cooperation along with our distinctive diets, life histories, and livelihoods. The presence on the African savannah of large mammals vulnerable to attack by cognitively advanced predators must have given substantial advantages to the members of groups that developed means of coordinating the hunt and sharing its sporadically acquired prey. Correspondingly, groups that had learned how to cooperate in these ways would have benefited from preferentially targeting large animals, as opposed to food acquired in smaller packages, and thereby enlarging the place of hunted meat in their diet. Bruce Winterhalder and Eric Alden Smith (1992):60 write:

> ...only with the evolution of reciprocity or exchange-based food transfers did it become economical for individual hunters to target large game. The effective value of a large mammal to a lone forager ...probably was not great enough to justify the cost of attempting to pursue and capture it, ...However, once effective systems of reciprocity or exchange augment the effective value of very large packages to the hunter, such prey items would be more likely to enter the optimal diet.

We think it likely that the distinctive aspects of the human livelihood thus coevolved with the distinctive aspects of our social behavior, most notably cooperation.

Two approaches inspired by standard biological models, have constituted the workhorses of our explanation, multi-level selection and gene-culture coevolution. Could it be that altruistic cooperation became common
among humans in the absence of these two processes? We think it unlikely. By contrast, explanations of the emergence and proliferation of cooperative behaviors based on gene-culture coevolution and multi-level selection are quite plausible. First, the models and simulations of our evolutionary past presented in the previous chapters provide strong evidence that in the relevant evolutionary environments, selective pressures based on the positive assortment of behaviors arising from the group-structured nature of human populations could have been a significant influence on human evolution. Second, we have also demonstrated the important contribution to the evolution of social preferences that could have been accomplished by the cultural transmission of empirically well-documented behaviors such as the internalization of norms, within-group leveling, and between-group hostility. Third, the nature of preferences revealed in behavioral experiments and in other observations of human behavior is consistent with the view that genuine altruism, a willingness to sacrifice one’s own interest to help others, including those who are not family members, and not simply in return for anticipated reciprocation in the future, provides the proximate explanation of much of human cooperation. These altruistic preferences seem unlikely to have evolved by kin selection or by means of the various models in which helping others is just self-interest with a long time horizon. Rather, these other-regarding group-beneficial social preferences are the most likely psychological consequence of the gene-culture coevolutionary and multi-level selection processes we have described.

13.2 The Future of Cooperation

Conclusive evidence about the origins of human cooperation will remain elusive given the paucity of the empirical record and the complexity of the dynamical processes involved. As in many problems of historical explanation, perhaps the best that one can hope for is a plausible explanation consistent with the known facts. This is what we have attempted to provide.

The challenge of explaining the origins of human cooperation has led us to the study of the social and environmental conditions of life of mobile foraging bands and other stateless small scale societies that arguably made up most of human society for most of the history of anatomically modern humans. The same quest has made non-cooperative game theory (which assumes the absence of enforceable contracts) an essential tool. But as Ostrom (1990), Taylor (1996), and other authors have pointed out, most forms of
contemporary cooperation are supported by incentives and sanctions based on a mixture of multilateral peer interactions and third party enforcement, often accomplished by the modern nation state.

It would thus be wise to resist drawing strong conclusions about cooperation in the 21st century solely on the basis of our thinking about the origins of cooperation in the Late Pleistocene. One may doubt, for example, that that lethal intergroup conflict today contributes to the altruism, civic mindedness or other social preferences that could underpin the more cosmopolitan forms of cooperation required to address global challenges such as climate change and epidemics.

But the fundamental challenges of social living and sustaining a livelihood faced by our distant ancestors are in many respects not fundamentally different from those we face today. Modern states and global markets have provided conditions for mutualistic cooperation among strangers on a massive scale. But altruistic cooperation remains an essential requirement of economic and social life.

The reason is that neither private contract or governmental fiat singly or in combination provide an adequate basis for the governance of modern societies. Social interactions in modern economies are typically at best these quasi-contractual. Some aspects of what is being transacted are regulated by complete and readily-enforceable contracts, while others are not. Transactions concerning credit, employment, information, and other goods and services where quality is difficult to monitor provide examples of quasi-contractual exchanges.

Where contracting is absent or incomplete, the logic of Adam Smith’s invisible hand no longer holds. Decentralized market fail to implement efficient allocations. But governments typically lack the information, and often the motivation, necessary to provide adequate governance where markets fail or are absent.

We now know from laboratory experiments that subjects in market-like situations with complete contracts tend to behave like the *Homo economicus* of Adam Smith of *The Wealth of Nations*, but their behavior when their contracts are not complete resembles more the virtuous citizens of the Adam Smith of *The Theory of Moral Sentiments*. Thus, where the invisible hand fails, the handshake may succeed. Kenneth Arrow wrote (1971 p. 22)

> In the absence of trust...opportunities for mutually beneficial cooperation would have to be foregone...norms of social be-
behavior, including ethical and moral codes [may be] . . . reactions of society to compensate for market failures.

Thus, social preferences such as a concern for the well-being of others and for fair procedures remain essential to sustaining society and enhancing the quality of life.

In a world increasingly connected not just by trade in goods but also by the exchange of violence, information, viruses, and emissions, the importance of social preferences in underwriting human cooperation, even survival, may now be greater even than among that small group of foragers that began the exodus from Africa 55,000 years ago to spread this particular cooperative species to the far corners of the world.
A1 Agent-Based Models

Agent-based modeling is a tool for analyzing certain types of complex dynamical systems. The actors in these systems are individual agents who share many characteristics, but differ on key characteristics that affect their relative reproductive success. The agents operate semi-autonomously, but are linked through a network of structured interactions. The individual characteristics in a complex system evolve through a process of replication, mutation, and selection of relatively successful individuals. Such dynamics are recursive, meaning that changes in one period become the basis for changes in future periods, and are non-linear, with the implication that they are generally incapable of being expressed as closed-form analytical solutions to sets of equations.

For instance, the individuals may be competing for jobs, or males and females who are interested in finding a good mate, or predators trying to catch prey and prey who are trying to avoid being caught. Suppose the individuals and their environment undergo mutation and structural change, so that the resulting system is sufficiently complex that it has “emergent properties” that cannot be determined simply by aggregating individual interactions, and the mathematics of the system is too complicated to solve in analytical form. Finally, suppose that each individual history of interaction affects his play, and individuals continually adapt by taking on the behaviors of other individuals who have done especially well, in terms of fitness, material gain or some other standard, in the course of time. In such a situation, agent-based modeling is an appropriate tool for investigation the dynamics of the interactive system.

Agent-based modeling is widely used in the natural and behavioral sciences, and the growth in its use has been extremely rapid since the advent of the powerful desktop computer. Such modeling (often called ‘simulation’) lies outside the two standard methods of gaining scientific knowledge: deduction and induction. Deduction means proving theorems, that is, showing
that certain mathematical conclusions follow from certain axioms (e.g., the Pythagorean Theorem follows from the laws of algebra). Induction means finding lots of evidence and drawing conclusions (e.g., all swans are white). Agent-based modeling is like deduction in that it starts with a rigorously specified computer program, but it is like induction in that it treats the operation of the program as a set of data points from which generalizations can be made. In particular, if a complex system has emergent properties, these can be ascertained by exhibiting an agent-based model in which these properties are seen and persist over many simulations.

We use agent-based models to study the dynamics of hypothetical populations that are structured to evolve in ways that we think represent long-term human evolution. Using this method, we can generate literally thousands of artificial histories allowing us to investigate the likely effects of differences in the benefits and costs of cooperative activities, the frequency of group conflict, the structure of a group’s socialization practices and other influences.

While agent-based modeling is important because many dynamic strategic settings are too complicated to admit standard mathematical analysis, it should not be thought that agent-based models are necessarily second-best. The assumptions made to permit explicit analytical solutions are often sufficiently unrealistic (e.g., continuous time, infinite numbers of individuals) that the agent-based model behaves more like the situations we are trying to model than does a tractable analytical model (Durrett and Levin 1994). Nevertheless, agent-based models require extensive experimentation to ensure that results accurately reflect the properties of the system.

The classical example of an agent-based model is Thomas Schelling’s (1978) study of residential segregation. Schelling showed that if families like having at least one neighbor of the same ethnic background, but otherwise prefer diversity to uniformity in neighbors, in the long run, family relocation is likely to lead to a high degree of segregation. In this case, segregation is an emergent property of the system because no individual family wanted this or sought it, and it was not predicted as the logical consequence of the system’s initial conditions using a set of differential equations. We can study residential segregation using Schelling’s idea by varying the number of ethnic groups, varying the intensity of preference for uniformity and diversity, and by varying the rules of neighborhood formation and population movement. For instance, Schelling predicted the “tipping” phenomenon often found in residential dynamics: when the fraction of res-
idents of a minority group in a community reaches a certain “tipping point,” the community more or less rapidly shifts towards that group becoming a majority.

An evolutionary agent-based model has, in addition to a stage game representing the interaction of agents, has a replication phase, in which individuals replicate according to some updating rule, often in proportion to their average success in the stage game. As a result, old individuals disappear, and their offspring inherit their behavior, perhaps with some mutation. In an evolutionary agent-based model, the more successful strategies are permitted to increase in frequency at the expense of the less successful. Thus, evolutionary agent-based models possess the three main characteristics of Darwinian evolution: replication, mutation, and selection according to fitness.

Among the first behavioral science applications of evolutionary agent-based modeling was Joshua Epstein and Robert Axtell’s Growing Artificial Societies (1997), which produced a wide variety of emergent behaviors from simple rules governing individual behavior inscribed in a “chromosome” that could evolve over time much as DNA, the computer genes being used to represent different strategic behaviors. They modeled migration, environmental externalities, conflict, and even disease transmission in a population of individuals competing for food. More recently, agent-based models of political competition, stock markets dynamics and investor strategies, as well as models of parasitism and disease transmission have become standard components of behavioral research (Miller and Page 2007, Epstein 2007). For a useful overview of the application of such models to various fields in the natural and behavioral sciences, see Tesfatsion and Judd (2006).

Figure A1 shows the programming structure of a typical evolutionary agent-based model. In the figure, “Game Parameters” refer to the specifics of the stage game being simulated, including the payoffs, the probabilities with which various events occur, and the like. The “Number of Generations” specifies how many rounds of replication you want to take place. This may be as small as 10 or as large as 10,000,000. The “Number of Rounds/Generation” refers to the speed of play as compared to the speed of replication. By the Law of Large Numbers, the more rounds per generation, the more accurately the actual success of individuals reflects the expected payoff of the strategies they represent. “Group All Individuals” captures the matching phase of the simulation, as well as the notion that a strategy in a game reflects a social practice occurring in the population,
rather than the carefully thought-out optimizing strategy of classical game theory. Note that in some situations, we will want some structure to this stage of the model. For instance, “neighbors” may meet more frequently than “strangers,” or individuals who play similar strategies may meet more frequently than individuals with different strategies.

The “Individuals Replicate” box is worth a diagram of its own, which we present in Figure A2. First we set various parameters, including the rate of mutation of new individuals and the extinction rate of old individuals. We then eliminate the appropriate number of unsuccessful individuals, and make an equal number of copies of the high success individuals, giving them the same strategies as their replicators, except that we allow some mutation. Success may be measured by fitness (offspring surviving to reproductive age) or by extent to which an individual’s behavior is copied (adopted through social learning) by others.
Game Parameters
Number of Individuals ($N$)
Number of Generations ($G$)
Number of Rounds/Generation ($K$)

Create $N$ Individuals

$g = 0$

$k = 0$

Group All Individuals

Individuals in each Group Play Game

Update Individual Payoffs

$k \rightarrow k + 1$

$k < K$? Yes

Individuals Replicate

$g \rightarrow g + 1$

$g < G$? Yes

Done

No

Figure A1. Structure of an Evolutionary Agent-Based Simulation
Suppose we have a number of individuals $i = 1, \ldots, n$ involved in a social interaction. Suppose each individual $i$ has a set $S_i$ of possible actions available to him, and each individual chooses an action $s_i \in S_i$ independently, and to each such action profile $(s_1, \ldots, s_n)$, each individual $i$ receives a payoff $\pi_i(s_1, \ldots, s_n)$. We call the various actions $s_i$ pure strategies and we call this social situation a game in strategic form. This definition extends readily to mixed strategies, where players use probability distributions over their pure strategies, but we will not need this extension in this book.
An example of a game in strategic form is the prisoner’s dilemma (§3.2), in which each player $i = 1, 2$ has pure strategy set $S_i = \{H, N\}$, and 
\[ \pi_1(H, H) = b - c, \pi_1(H, N) = -c, \pi_1(N, H) = b, \text{ and } \pi_1(N, N) = 0. \]
The payoffs to player 2 are 
\[ \pi_2(H, H) = b - c, \pi_2(H, N) = b, \pi_1(N, H) = -c, \text{ and } \pi_2(N, N) = 0. \]

Suppose players choose strategies $(s_1, \ldots, s_n)$. We say $s_i$ is a best response to the remaining $n - 1$ strategies of the other players if there is no strategy for player $i$ that can give player $i$ a payoff greater than $\pi_i(s_1, \ldots, s_n)$. We say $(s_1, \ldots, s_n)$ is a Nash equilibrium if every player’s choice is a best response to the choices of the other players. There are many cases in which we do not expect players to choose a Nash equilibrium, for example the repeated games described in Chapter 4, but in many games described in this book, it is reasonable to expect individuals to play a Nash equilibrium.

In some games, including the prisoner’s dilemma described above, and its extension to an $n$-player game, the public goods game (§3.3), a player may have a strategy that offers higher payoff than any other strategy, no matter what the other players do. Such a strategy is called dominant. A player who has a dominant strategy will play it in any Nash equilibrium, because such a strategy is a best response to any configuration of strategies of the other players. In the one-shot prisoner’s dilemma and public goods games, for instance, defecting is a dominant strategy for all players.

In many games, players move more than one time and do not play simultaneously. We represent such games as extensive form games, as in Figure 8.1. In this case we use a game tree consisting of nodes connected by branches. Each node represents a point in the game where a particular player gets to move, and each branch emanating from a node represents the various actions that the player has when choosing at that node. At the end of the game tree are terminal nodes where the payoffs to the players are exhibited. For a more complete treatment of extensive form games, consult Gintis (2009b).

We can define a Nash equilibrium for an extensive form game in a manner parallel to the definition in the previous section. Consider a choice of an action at each node of the extensive form game. We say the resulting profile of choices is a Nash equilibrium in no player can gain by changing his choice at any node where he gets to choose.

Suppose we start with a game $G$ as in the previous section, and repeat the game indefinitely, with a probability $1 - \delta > 0$ of terminating the pro-
cess at the end of each period (§3.2). We then call $G$ the stage game, of the repeated game. The payoff to the repeated game is just the sum of the payoffs to the various stages. Note that the payoff is finite, because the process terminates in a finite number of periods with probability one. Note also that we can include a time discount factor in $\delta$, as explained in Chapter 3.

The most important fact about the repeated game based on index-term stage game $G$ is that it can support cooperative equilibria in situations where $G$ cannot. Consider, for instance, the prisoner’s dilemma (§3.2). The only Nash equilibrium of the stage game has both players defecting, hence earning payoff 0. But, suppose the players follow the strategy in the repeated game of cooperating until the first player defects, and then defecting forever. We call this the Grim Strategy. By following the Grim Strategy, the players earn $(b - c)/(1 - \delta)$, as analyzed in §6.1. The gain from defecting right away is $b$, so following the Grim Strategy is a Nash equilibrium as long as $(b - c)/(1 - \delta) > b$, which reduces to $\delta b > c$.

An evolutionary game is a repeated game, as in the last section. However, instead of $n$ players, there are $n$ populations of players, and in each period, many stage games $G$ are played, each consisting of one player drawn randomly from each population. Moreover, as in agent-based modeling, we assume players each play a fixed strategy rather than a best response. A running total of player’s scores is kept, and periodically, the high-scoring players have their strategies copied by the low-scoring players. Moreover, we may add a bit of random mutation to the mix, so that once in a while a low-scoring agent chooses a random strategy rather than imitating a high-scoring individual.

The idea of an evolutionary game is due to the biologist Maynard Smith (1982), who applied it to animals whose limited intelligence does not allow us to assume that they play best responses. However, the concept applies just as well to humans who, while intelligent, often lack the information to choose a best response, or who are not sufficiently coordinated to choose one among the variety of best responses available to them.

The most natural dynamic to apply to an evolutionary game is the replicator dynamic, which we describe below. Indeed, it can be shown that every equilibrium of an evolutionary game under the replicator dynamic is a Nash equilibrium of the stage game (Nachbar 1990). This shows that the Nash equilibrium criterion remains powerful even without assuming that players are rational (i.e., that they choose best responses) or coordinated.
However, in many cases, Nash equilibria of the stage game are not stable equilibria of the corresponding evolutionary game. For this reason, Maynard Smith developed the stronger notion of an evolutionarily stable strategy (ESS) for the case of a two-player stage game $G$ in which a single population of individuals play against themselves, as in the prisoner’s dilemma. A strategy is an ESS if a whole population using that strategy cannot be invaded by a small group playing any other strategy defined in $G$.

It is easy to show that an ESS is always a Nash equilibrium, but the converse is false. Indeed, an ESS is always a stable equilibrium of the replicator dynamic (Gintis 2009b).

We here derive the replicator dynamic for an evolutionary game by imitation of cultural (phenotypic) traits. The derivation for genetic evolution is similar, but somewhat simpler to derive. Consider a population of individuals who play a game in which each player follows one of $n$ pure strategies $s_i$ for $i = 1, \ldots, n$. The play is repeated in periods $t = 1, 2, \ldots$. Let $p_i^t$ be the fraction of players playing $s_i$ in period $t$, and suppose the payoff to $s_i$ is $\pi_i^t = \pi_i(p^t)$, where $p = (p_1, \ldots, p_n)$. We look at a given time $t$, and number the strategies so that $\pi_1^t \leq \pi_2^t \leq \ldots \leq \pi_n^t$.

Suppose in every time period $dt$, each individual with probability $\alpha dt > 0$ learns the payoff to another randomly chosen other individual and changes to the other’s strategy if he perceives that the other’s payoff is higher. However, information concerning the difference in the expected payoffs of the two strategies is imperfect, so the larger the difference in the payoffs, the more likely the individual is to perceive it, and change. Specifically, we assume the probability $p^t_{ij}$ that an individual using $s_i$ will shift to $s_j$ is given by

$$p^t_{ij} = \begin{cases} 
\beta(\pi_j^t - \pi_i^t) & \text{for } \pi_j^t > \pi_i^t \\
0 & \text{for } \pi_j^t \leq \pi_i^t
\end{cases}$$

where $\beta$ is sufficiently small that $p_{ij} \leq 1$ holds for all $i, j$. The expected fraction $E p_{i}^{t+dt}$ of the population using $s_i$ in period $t + dt$ is then given by

$$E p_{i}^{t+dt} = p_i^t - \alpha dt \ p_i^t \ \sum_{j=i+1}^{n} p_j^t \beta(\pi_j^t - \pi_i^t) + \sum_{j=1}^{i} \alpha dt \ p_j^t p_i^t \beta(\pi_j^t - \pi_i^t)$$

$$= p_i^t + \alpha dt \ p_i^t \ \sum_{j=1}^{n} p_j^t \beta(\pi_i^t - \pi_j^t)$$

$$= p_i^t + \alpha dt \ p_i^t \beta(\pi_i^t - \pi^t).$$
where $\bar{\pi}^t = \pi_1^t p_1^t + \ldots + \pi_n^t p_n^t$ is the average return for the whole population. If the population is large, we can replace $E p_i^{t+d}t$ by $p_i^{t+d}t$. Subtracting $p_i^{t}t$ from both sides, dividing by $d t$, and taking the limit as $d t \to 0$, we get

$$\dot{p}_i^t = \alpha \beta p_i^t (\pi_i^t - \bar{\pi}^t), \quad \text{for } i = 1, \ldots, n,$$

which is called the replicator dynamic. Because the constant factor $\alpha \beta$ merely changes the rate of adjustment to stationarity but leaves the stability properties and trajectories of the dynamical system unchanged, we often simply assume $\alpha \beta = 1$.

Several points are worth making concerning the replicator dynamic. First, under the replicator dynamic, the frequency of a strategy increases exactly when it has above average payoff. Second, if we add up all the equations, we get $\sum_i \dot{p}_i^t = 0$, so if $\sum_i p_i^t = 1$ at one point in time, this remains true forever. Moreover, while a particular replicator can become extinct, a replicator that is not represented in the population at one point in time will never be represented in the population at any later point in time. So, replicator dynamics deal poorly with mutation and innovation. A more general system adds a term to the replicator equation expressing the spontaneous emergence of replicators.


### A3 Continuation probability and Time Discount Factor

Suppose a game is played in periods $t = 1, 2, \ldots$ such that in each period the game continues for one more period with probability $\delta$, where $0 < \delta < 1$. Suppose a player has payoff $\pi$ in each period that the game is played, but has a time discount factor $d$, with $0 < d < 1$, such that a payoff $\pi$ in period $t$ is worth $d^t \pi$ at the start of the game. Then the expected value $v$ of the game at the beginning of the first period satisfied the equation

$$v = \pi + \delta dv,$$

so, letting $\delta^* = \delta d$, we have

$$v = \frac{\pi}{1 - \delta^*}.$$  

(A2)
Because the continuation probability $\delta$ and the time discount factor $d$ enter multiplicatively in (A2), we can interpret $\delta^*$ as an arbitrary combination of continuation probability and time discount factor.

### A4 Alternatives to the Standing Model

To investigate the sensitivity of the standing model of §3.4 to informational requirements, suppose that with probability $q$ an individual knows the standing of his partner, and with probability $1 - q$ he has no information concerning the partner’s status. Suppose that an individual following the cooperative strategy who is in good standing cooperates unless he knows his new partner is in bad standing.

Using this model, and assuming an execution error rate of $\epsilon \geq 0$, we find that the information requirements for indirect reciprocity being an evolutionary stable strategy are that the minimum feasible $q$ satisfies

$$q_{\text{min}} > \frac{c}{b} \left(1 + \epsilon \left(1 - \frac{c}{b}\right)\right)$$

(Gintis 2007b). This requirement is demanding except perhaps in the smallest groups, and is not likely to be met in most real-world conditions.

The indirect reciprocity model can be extended to groups of size $n > 2$ in the obvious fashion, where an individual is in good standing if he cooperated in the previous period, or when he defected while being in goods standing in the previous period in a group with at least one member in bad standing. But, the informational requirements clearly become prohibitive under normal conditions for larger groups.

Nowak and Sigmund (1998b) addressed the problem of the excessive informational requirements of the standing strategy by investigating an indirect reciprocity strategy which they term image scoring. Nowak and Sigmund show that the strategy of cooperating with others who have cooperated in the past, independent of the standing of the Cooperator’s partner, is stable against invasion by Defectors, and weakly stable against invasion by Cooperators once Defectors are eliminated from the population (i.e., when there are no Defectors in the population, Cooperators and image scorers have equal payoffs). However, Panchanathan and Boyd (2003) showed that if execution errors, however small, are introduced into the model, cooperation becomes unstable, and universal defect becomes the only stable equilibrium of the system. Nowak and Sigmund (1998a) use an agent-based
model (§A1) to analyze the stability of a somewhat more complex image scoring strategy, but (Leimar and Hammerstein 2001) show that a strategy not considered by Nowak and Sigmund, that of when one’s own image score is low, independent of the image score of the recipient, can invade the image scoring strategy. The weakness of image-scoring lies in the fact that in an image-scoring model, there is no incentive for a self-regarding individual to care about the status of the potential recipient when deciding whether or not to help.

However, the image-scoring model is considerably less information-demanding that the standard indirect reciprocity model, so it is useful to investigate conditions under which it could survive whereas a standing model would not. Thus, Brandt and Sigmund (2004) show that the robustness of image scoring dramatically increases when there are perceptual errors in judging whether an individual has defected. This is because in such circumstances, there is extreme lack of agreement as to who is in good standing, and hence when a defection is to be considered a reason for falling into bad standing. Because the image scoring process does not take into account the status of the recipient of help, it makes far fewer errors where perceptual errors are frequent.

Brandt and Sigmund (2005) investigate an alternative scenario in which each player $i$ in the indirect reciprocity game knows the status of $q_i$ other players, and the average level of information $q = \sum_i q_i / n$ increase from period to period. For instance, when an individual helps another individual, he may add his new partner to his “circle of friends,” and may have a way of retaining information concerning his friends’ status in future periods of play. Brandt and Sigmund (2005) show that under these conditions, image scoring can resist invasion by Defectors.

### A5 The Prisoner’s Dilemma with Public and Private Signals

Suppose in (§4.3) that the signals are public, and the continuation probability $\delta$ and error rate $\epsilon$ are such that cooperation can be sustained with the OBSTD strategies, and if $\pi_r$ is the expected payoff to playing OBSTD, we must have

$$\pi_r = 5 + (1 - \epsilon)^2 \delta \pi_r$$

(A3)
Here, the first term represents the payoff in the first period, and the second term reflects the fact that with probability \((1 - \epsilon)^2\) neither player receives a defect signal in error, so the game continues another period with probability \((1 - \epsilon)^2 \delta\). Thus we have, in the case of public signals,

\[
\rho^*_r = \frac{5}{1 - \delta (1 - \epsilon)^2}, \tag{A4}
\]

and that playing OBSTD beats intentionally defecting, which has expected payoff 8, provided

\[
\delta \geq \frac{3}{8(1 - \epsilon)^2}.
\]

If the error signals is private, but players still use OBSTD, then \((A3)\) becomes

\[
\rho_r = 5 + (1 - \epsilon)^2 \delta \rho_r - 3 \delta \epsilon (1 - \epsilon) + 8 \delta \epsilon (1 - \epsilon) + \epsilon^2 (0). \tag{A5}
\]

Here, the first two terms are as before, the third term represents the case that Bob cooperates but Alice receives the defect signal. Bob thus cooperates in next period when Alice defects, because he does not know that Alice received the defect signal, so Bob receives -3, and both defect in every period after that. The fourth term reflects the case that Alice cooperates but Bob receives the defect signal. Bob thus defects in the next period but Alice cooperates, so Bob’s payoff is 8. The expected value of this game is, after some algebra, is given by

\[
\rho_r = \frac{5 (1 + \delta \epsilon (1 - \epsilon))}{1 - \delta (1 - \epsilon)^2}, \tag{A6}
\]

which is close to from \((A4)\). For instance, if \(\delta = 0.9\) and \(\epsilon = 0.05\), \(\rho_r\) from \((A6)\) is is about 1.5% greater than \(\rho_r\) from \((A4)\).

To see that with private signals, Alice’s TBSTD strategy has higher payoff that OBSTD when played against Bob’s OBSTD strategy, we must consider several possible states of the game. Let \(gg\) be the state where both players received the good (cooperate) signal in the previous period. In this case each player has payoff 5 for cooperating. Both players in the \(gg\) state cooperate, so with probability \((1 - \epsilon)^2\) the next state will also be \(gg\). With probability \(\epsilon (1 - \epsilon)\), however, Alice will receive the bad (defect) signal while Bob receives the good signal. We call this state \(gb\). Also with probability \(\epsilon (1 - \epsilon)\) Bob will receive the bad (defect) signal while Alice
receives the good signal. We call this state bg. Finally, with probability $\epsilon^2$ both players receive the bad signal. We call this state bb. If we write $\pi_s$ for the expected payoff of the game starting in state $s$, we then have the equation

$$\pi_{gg} = 5 + \delta[(1 - \epsilon)^2\pi_{gg} + \epsilon(1 - \epsilon)(\pi_{bg} + \pi_{gb}) + \epsilon^2\pi_{bb}].$$

In state gb, Alice ignores the defect signal so both players cooperated, with payoff 5. Then with probability $(1 - \epsilon)^2$ both receive the good signal, so the next state will be gg. With probability $\epsilon(1 - \epsilon)$ Alice receives a second bad signal in a row but Bob receives a good signal. We call this state gbb. With probability $\epsilon(1 - \epsilon)$ Alice receives a good signal but Bob receives a bad signal. This is just state bg. Finally, with probability $\epsilon^2$ both players receive the bad signal, which is then the second in a row for Alice. We call this state bbb. We then have the equation

$$\pi_{gb} = 5 + \delta[(1 - \epsilon)^2\pi_{gg} + \epsilon(1 - \epsilon)(\pi_{gbb} + \pi_{bg}) + \epsilon^2\pi_{bbb}].$$

We also have

$$\pi_{bg} = -3(1 + \delta)$$
$$\pi_{bb} = -3$$
$$\pi_{gbb} = 5 + 8\delta)$$
$$\pi_{bbb} = 0.$$  

The first equation follows from the fact that Bob defects unconditionally forever, while Alice ignores the first bad signal. The second equation is the same except Alice only cooperates once. In the third equation, Bob has actually cooperated twice but Alice received two bad signals, so she defects on a cooperating Bob in the second round. The final equation is obvious.

If we solve these six equations simultaneously, we will get a value $\pi_r = \pi_{gg}$ for the original game. The resulting equation for $\pi_r$ is long and complex, so we won’t present it here. Comparing this payoff with (A6), we find that unless $\epsilon$ is very large, Alice’s TBSTD strategy is better than OBSTD against OBSTD.

To find a Nash equilibrium in the private signal case, following Sekiguchi (1997) and Bhaskar and Obara (2002), suppose players cooperate with probability less than one on the first round and then in following rounds, they use a trigger strategy in which the first time a player receives a
defect signal, he defects forever. To explain the first period, and assess the
efficiency properties of the equilibrium, assume that from the second period
on, Bob and Alice play the prisoner’s dilemma depicted in §4.1, which they
play repeatedly with discount factor $\delta$. Now suppose Bob cooperates on the
first round with probability $p_o$. Then if Alice cooperates on the first round,
her expected payoff is

$$\pi_c = 5p_o - 3(1 - p_o) + 3(1 - \epsilon)\delta \pi_r$$

and if she defects, her payoff is

$$\pi_d = 8p_o.$$  

To get Alice to cooperate, we must have $\pi_c \geq \pi_d$. Assuming equality, we
get the equilibrium probability $p_o = p^*$ given by

$$p^* = \frac{3(1 - \delta(1 - \epsilon)^2)}{5\delta(1 - \epsilon)}.$$  \hspace{1cm} (A7)

This $p^*$ is a probability, so it must lie between zero and one. We certainly
have $p^* > 0$, and we will have $p^* < 1$ provided $\delta$ is near unity and $\epsilon < 0.8$.
Hence for plausible values of the parameters, a private signal equilibrium
exists in this case. The problem is that this equilibrium is extremely ineffi-
cient. For instance, with $\delta = 0.90$ and $\epsilon = 0.05$, the efficiency is about
3%. By contrast, as we have seen, in the public signal case with the same
parameters, efficiency is over 50%.

### A6 Student and Nonstudent Experimental Subjects

Are students who volunteer for experiments more pro-social than the gen-
eral public? Bellemare et al. (2008) in an inequity aversion study, write that
“...extending the subject pool from students only to a more representative
population [of Dutch citizens]... generates a distribution with much greater
levels of inequity aversion.” In other words, students are considerably less
inequity averse than the more general Dutch sample studied.

Baran et al. (2009) found that an individual’s sensitivity to social pres-
sure, as measured by the Crown-Marlowe Social Desirability Scale (Crowne
and Marlowe 1960) influenced prosocial in a natural setting, that of contribu-
tions to the University of Chicago Graduate School of Business, but not
Appendices 271

in a laboratory setting, a trust game in which the measure of prosocial behavior was the fraction of money returned by the trustee.

Carpenter et al. (2005) used a dictator game showing that Middlebury College and Kansas City Community College students contributed less than employees at a Kansas City warehouse. Kansas City Community College students offered more in an ultimatum game than warehouse workers, who in turn offered more than Middlebury College students. Workers offered the same amount in the dictator game and the ultimatum game, while students gave much less in the dictator game, consistent with the view that in the ultimatum game, workers were not making strategic offers, while students were.

Burks et al. (in press) used a sequential prisoner’s dilemma game with subjects drawn from a sample of students on the one hand and professional bicycle messengers on the other, in Switzerland and the United States. The bicycle messengers exhibited more cooperative behavior than the students.

List (2004) studied contributions in a public goods game, finding that the behavior of participants at a sports-card show closely approximated that of students in laboratory experiments such as Fehr and Gächter (2000a), Isaac et al. (1994), and Andreoni (1988). Participants aged 49 years and older cooperated more than college-age students, while middle-aged subjects approximated the behavior of college-age subjects. Controlling for income, older Florida residents also contributed more in a university fund-raising appeal. In the television game show Friend or Foe, a prisoner’s dilemma game for big stakes, older players were more likely to cooperate than younger.

Carpenter et al. (2008) used a dictator game with subject-named charity recipients, finding that students gave twenty-five percent less than non-students. Students also gave significantly less, controlling for age and other demographics, and older respondents gave more, controlling for student status. Forty-eight percent of non-students contributed the entire endowment; only sixteen percent of students did.

Cardenas (2005) used a common pool resource game, finding that a sample of students from Columbia were less cooperative, extracting significantly more from the common pool, than a sample of Colombian villagers.

Cleave et al. (2010) found that students who volunteered for an experiment were as trustworthy but less trusting than other students from the population from which they were drawn, but who were administered the trust game as a captive audience.
Falk et al. (2010) found that Swiss students who exhibited strong pro-social behavior in an unrelated field donation, that of contribution to a social fund, were not more likely to participate in experiments than students who did not contribute the social fund. The experimenters also found that students and the general population were equally trusting in a trust game, but students trustees returned a smaller fraction of what was given to them by trusters.

### A7 Dynamical Systems

We will explain two major types of dynamical systems, a continuous time system using differential equations, and a discrete time system using Markov chains. The material presented below is a condensed version of the presentation in Gintis (2009b), to which the reader should refer for additional material and references.

For ease of exposition, suppose our continuous time system has two dimensions, represented by variables $x$ and $y$, the horizontal and vertical axes of the Cartesian plane. An example is the coevolution of the fraction of altruists and of parochialists in the population depicted in Figure 9.1 and Figure 9.2. The path of the system in time is represented by a pair of functions $x(t)$ and $y(t)$. We assume that the rate of change of the system in the $x$- and $y$-directions are functions of the position of the system alone, so we can write

\[
\frac{dx(t)}{dt} = f(x, y) \tag{A8}
\]

\[
\frac{dy(t)}{dt} = g(x, y). \tag{A9}
\]

There is a theorem that guarantees that, given an initial time $t = t_0$, if $f$ and $g$ are reasonably well-behaved, there is a unique solution to (A8) through $t = t_0$. This is called the path of the dynamical system.

An equilibrium of the dynamical system (A8), also called a critical point or fixed point, or stationary point, is a value $(x^*, y^*)$ such that $f(x^*, y^*) = g(x^*, y^*) = 0$. Note that at an equilibrium, $dx/dt = dy/dt = 0$, so the dynamical system remains forever at $(x^*, y^*)$ once it reaches there. Under what conditions does a dynamical system move towards an equilibrium?
Suppose that, starting from a point \((x_0, y_0)\), the path of the dynamical system (A8) approaches \((x^*, y^*)\) as \(t \to \infty\); i.e., \(\lim_{t \to \infty} x(t) = x^*\) and \(\lim_{t \to \infty} y(t) = y^*\). Then we say \((x_0, y_0)\) is in the basin of attraction of the equilibrium \((x^*, y^*)\). If the basin of attraction of an equilibrium is two-dimensional, it must surround the equilibrium, and we say the equilibrium is asymptotically stable, or simply stable. If the equilibrium is not stable, it may be either neutrally stable or unstable. The equilibrium is unstable if, no matter how close a path starts near (but not at) the equilibrium, there is a positive distance \(d\) from the equilibrium such that the path eventually is farther than \(d\) from the equilibrium, and never gets closer than \(d\) thereafter. If the path is neither stable nor unstable, we say it is neutrally stable. In the neighborhood of a neutrally stable equilibrium, paths neither escape nor converge to the equilibrium. They are therefore themselves either equilibria, or they trace out paths around the equilibrium.

Very few dynamical systems, even simple ones in two dimensions, can be solved analytically, so the paths \(x(t)\) and \(y(t)\) cannot be written in closed form. Nevertheless, there are well-developed methods for determining when an equilibrium is stable, unstable, or neutrally stable, using tools from algebra and calculus.

A finite Markov chain is a dynamical system that can be in any of \(n\) states \(s_1, \ldots, s_n\), and if the system is in state \(i\) in time period \(t\), it will be in state \(j\) in time period \(t + 1\) with probability \(p_{ij}\). Of course, for this to make sense, we must have \(p_{ij} \geq 0\) for all \(i, j = 1, \ldots, n\), and \(\sum_{j=1}^{n} p_{ij} = 1\).

Statistical estimates of these probabilities, based on thousands of implementations of our model, for example, are the basis of our calculation the vector field in Figure figeightthree giving the movement of the population among the states indicating various frequencies of altruists and of parochials.

For instance, consider two urns, one filled with 10 red balls and the other with 10 white balls. In each time period, we choose one ball from each urn simultaneously and place each ball in the other urn. Let \(s_i\) be the state where the first urn has \(i\) red balls, so we start out in state \(s_{10}\). It is easy to write down the transition probabilities if you know a little probability theory, but we will leave this exercise to the reader. It is intuitively obvious that in the long run, there will be an average of 5 white balls in the first urn, and this will be true independent of how many of the 10 white balls were in this urn when we started. When a Markov chain has the property that the average fraction of time in each state in the long run is independent from the starting state, we say the system is ergodic, and we call the...
resulting long run distribution of probabilities the stationary distribution of the Markov chain. The Markov chain represented by our urn problem is thus an ergodic Markov chain. Figure 9.2 gives the stationary distribution of parochials and altruists while Figure 3.3 and Figure 3.4 give the mean levels of net cooperation in the stationary distribution of the agent-based models of cooperation, for varying group sizes and error rates.

We say the states of a Markov chain communicate if, for every pair of states, there is a positive probability of moving from the first state to the other and then back. If we write \( p_{ij}^{(k)} \) for the probability that the Markov chain moves from state \( i \) to state \( j \) in \( k \) periods, then states \( i \) and \( j \) communicate if \( p_{ij}^{(k)} > 0 \) and \( p_{ji}^{(m)} > 0 \) for some integers \( k \) and \( m \). Let \( i \) be a state of a Markov chain, and let \( T_i \) be the set of integers \( k \) such that \( p_{ii}^{(k)} > 0 \). If the greatest common divisor of the integers in \( T_i \) is greater than 1, we say the state \( i \) is periodic. For instance, consider a Markov chain consisting of 10 points equally spaced on a circle, numbered 1 to 10, and suppose the state can move either clockwise or counter-clockwise by one position. Then, every state has period 2, because the system can only return to a state in an even number of transitions. If a Markov chain has no periodic states, we say the Markov chain is aperiodic.

Using these definitions, we can state the most important property of finite Markov chains. Suppose every pair of states of a finite aperiodic Markov chain communicate. Then the chain is ergodic. Clearly, this applies to the above urn problem.

An agent-based model is a finite Markov chain because there are a finite number of agents, each can only be in a finite number of states, and if there are parameters in the model (e.g., the current weather), then these are in finite number. Moreover, a computer language can support only a finite number of real numbers without ‘overflow’, so even supposedly “real” numbers are really a finite range of integral fractions. The number of states in the resulting system can be huge, but it is a finite number. Moreover, while it may be quite impractical to calculate the probability of movement from one state to another, the probability itself is perfectly determinate. To ensure that the Markov chain represented by the agent-based model is ergodic, we always allow agents to remain in their current state with positive probability, and mutate to another state with positive probability, however small.

It follows that the ergodic theorem holds for all of the agent-based models used in this book. This means that the long-run behavior of the dynam-
ical systems generated by our agent-based models is independent from the particular initial parameters we have chosen for our illustrative runs. Of course, the ergodic theorem cannot tell us how long it will take to “erase” the effect of our initial conditions, and the number of periods involved could be truly astronomical. Thus, we take the question of how fast a system moves to its stationary distribution to be a very important one.

It may appear that the stationary distribution of an ergodic Markov chain somehow represents a small cluster of nearby “long-run average states,” but that is not the case in some Markov chains. For instance, consider the Markov chain with two states $s_1$ and $s_2$, where $p_{11} = 0.99$, $p_{12} = 0.01$, $p_{22} = \epsilon$ and $p_{21} = 1-\epsilon$. If $\epsilon$ is a small perturbation, this chain spends almost all its time in state 1, but it does spend a positive amount of time, about $\epsilon$ periods, on average, in state 2. We say a state is recurrent if it spends a positive fraction of time in this state in the stationary distribution, so both states in this example are recurrent. However, when it enters state 1, it stays there for almost 100 periods, on average, before moving to states 2, while in state 2, it rarely stays for more than one period. We say that a state is stochastically stable if, under a small perturbation of size $\epsilon > 0$, the fraction of time it spends in that state is bounded away from zero as $\epsilon \to 0$. Clearly, only state 1 is stochastically stable in this case. Young (1998) uses finite Markov processes to study the long term evolution of contracts and other institutions.

### A8 Altruism Defined

In Chapters 5 and 6 we provide two definitions of altruistic behavior, one concerning preferences (other-regarding or self-regarding) that motivate behavior (p. 15) and the other based on its fitness effects on the actor and others (p. 110).

Our biological definition of altruism refers to a large population composed of a large number of groups: the behaviors induced by an altruistic genotype increase the expected average fitness of members of the group of which the focal individual is a member, but the focal individual would have higher fitness were he not the bearer of the altruistic genotype. This is the definition adopted by Hamilton (1975), Grafen (1984), and others. To make this clear, following Kerr et al. (2004), suppose the expected fitness of altruists and non-altruists in a group of $m$ members with $j$ altruists are $w^A(j)$ and $w^N(j)$ respectively. Then our first condition, that the altruistic
behavior raises the expected average fitness of members of the group, is that

\[ P = \frac{jW^A(j + 1) + (m - j - 1)w^N(j)}{m} > \frac{jW^A(j) + (m - j)w^N(j)}{m} \]

and the second, that switching from an \( N \) to an \( A \) lowers the actor’s fitness, is

\[ w^A(j + 1) < w^N(j). \]  

These two conditions are respectively Conditions 4 and 1 of Kerr et al. (2004), which together imply their Condition 2.

There are two attractive features of this definition. First, it is the most stringent definition possible (other than definitions that preclude the evolution of altruism under any conditions). In the absence of positive assortment altruism cannot evolve under our definition. An alternative definition found in Haldane (1932), Cohen and Eshel (1976), Maynard Smith (1964) and Sober and Wilson (1998) replaces (A10) with the condition that \( A \)’s have lower fitness than \( N \)’s, or

\[ w^A(j) < w^N(j). \]  

This is less stringent because it does not preclude that \( W^A(j + 1) > w^N(j) \) so that an \( N \) would increase its fitness by switching to an \( A \). This would be the case if the altruistic act cost \( c \) and conferred a benefit \( b \) on a randomly selected member of the group, and \( b/m > c \). Then condition 3 would hold but condition 2 would not: by switching to \( A \) from \( N \), the chance that the benefit would accrue to the actor would more than offset the cost of the behavior. In this case altruism will evolve even under random assortment (Matessi and Jayakar 1976).

The second attractive feature is that the definition maps directly onto the two terms in the Price equation (A13). Recall that the \( \beta \)’s in the equation are partial derivatives of expected group average fitness with respect to (for \( \beta_G \) the fraction of altruists in the group and (for \( \beta_i \) the actor’s own type (\( N \) or \( A \)). Condition 1 above requires that \( \beta_G > 0 \) while condition 2 requires that \( \beta_i < 0 \), thus defining what we have called (in Chapter 3) the strong group selection problem.

Our preference-based definition is that a person acts altruistically by conferring benefits on other members of one’s group at a personal cost, meaning that the act would not be chosen by an individual with entirely self-regarding preferences. Strictly speaking, our definition does not require that
the average well-being of the group members increase, but we are generally interested in altruistic behaviors only when in some sense this is true. The definitions are identical if the states over which the individual is choosing are fitness levels, so that an entirely self-regarding individual would never act altruistically (if condition 2 above obtained) because given 2, \( N \) is the unique best response independently of \( j \). Because we also generally assume that \( b \), the benefit conferred on others, exceeds \( c \), the cost, condition 2 must obtain: an altruistic act raises the total benefit of the group. This formulation entails that the well-being of group members can be summed, if not in fitness terms, then in some other.

The attractive feature of this definition is that it allows altruistic behaviors to be motivated by preferences, thus allowing use of the term altruistic or other-regarding preferences. The alternative, to regard any preference-motivated behavior as selfish, is the equivalent in preference terms to the biological definition requiring that altruistic behaviors entail lower fitness in the population taken as a whole. The first requires that altruism be counter-preferential and hence irrational. The second precludes altruism being evolutionarily feasible.

### A9 The Price Equation

Suppose there are groups \( j = 1, \ldots, m \), and let \( q_j \) be the fraction of the population in group \( j \). Let \( w_j \) be the mean fitness of the members of group \( j \), so \( w = \sum_j q_j w_j \) is the mean fitness of the whole population. We use the term fitness here, but as the Price equation is applicable to any system in which traits are differentially replicated over time, we could have used the more general term: number of replicas in the next period. Groups grow from one period to the next in proportion to their relative fitness, so if \( q'_j \) is the fraction of the population in group \( j \) in the next period, then

\[
q'_j = q_j \frac{w_j}{w}.
\]

Suppose there is a trait with frequency \( p_j \) in group \( j \), so the frequency of the trait in the whole population is \( p = \sum_j q_j p_j \). If \( p'_j \) and \( p' \) are the mean fitness of members of group \( j \) and the frequency of the trait in group \( j \) in the next period, respectively, then \( p' = \sum_j q'_j p'_j \), so

\[
p' - p = \sum_j q'_j p'_j - \sum_j q_j p_j
\]
Now writing $\Delta p = p' - p$ and multiplying the above expression by $w$, this becomes

$$w \Delta p = \sum q_j (w_j - w) p_j + \sum q_j w_j \Delta p_j.$$  \hfill (A12)

Because $\sum q_j (w_j - w) p = 0$, from the definition of $w$, and noting that

$$\sum q_j (w_j - w)(p_j - p) \equiv \text{cov}[w_j, p_j]$$

and

$$\sum q_j w_j \Delta p_j \equiv \mathbb{E}[w_j \Delta p_j],$$

we can rewrite A12 as

$$w \Delta p = \text{cov}[w_j, p_j] + \mathbb{E}[w_j \Delta p_j],$$  \hfill (A13)

where the covariance and expectations are taken with respect to the population fractions $q_j$. We interpret (A13) as follows. The trait measured by $p$ will increase in frequency if the left hand side of the equation is positive. Because $w > 0$, this will occur only if the right hand side is positive. If the trait measured by $p$ is altruistic, as in the conventional application of the Price equation, then the first term $\text{cov}[w_j, p_j]$ will be positive, because groups with higher fractions of the trait will have supranormal average fitness of its members. But, the second term, $\mathbb{E}[w_j \Delta p_j]$, will be negative, because the altruists are disadvantaged within each group, so their within-group frequency declines, and $\Delta p_j$ will be negative for all $j$. The trait will then spread only if the between-group advantage of the high altruism groups is sufficient to overcome the within-group disadvantage of the altruists.

In Chapter 10 we apply the above equation to the case of Nonpunishing Cooperators, who may be considered within-group parasites rather than altruists, because they free ride on the punishment meted out by strong reciprocators against selfish individuals. In this case, the term $\text{cov}[w_j, p_j]$ may be negative, because groups with higher fractions of Cooperators may have been invaded by selfish individuals and hence have below average fitness, unless the fraction of Cooperators is very high. Within groups the situation is inverted. As long as there is a sufficient number of strong reciprocators and sufficiently few selfish individuals, the term $\mathbb{E}[w_j \Delta p_j]$ will be
positive, because the Cooperators are advantaged within each group. The reason is that they are never punished and do not bear the costs of punishing others. As a result, they are advantaged within each group, so their within-group frequency increases. The parasitic trait will then spread as long as the between-group disadvantage of the high Cooperator groups is insufficient to overcome the within-group advantage of the Cooperators. For very high or very low fractions of Cooperators in the population, of course this situation would not obtain. But the accounting of the within- and between-group effects in Figure 6.1 shows that for the simulations we have studied, our interpretation of the parasitic nature of Cooperators is indeed true.

In §6.1, we apply the Price equation to altruism in a social dilemma, where altruists supply a benefit $b$ to other members of the group at a cost $c$ to themselves. In this case, adding (6.1) over all $i$ in group $j$, of which there are $q_j N$, where $N$ is total population size, and dividing by $N$, we get

$$q_j w_j = q_j \beta_o + q_j p_j \beta_g + q_j p_j \beta_i.$$  \hfill (A14)

Adding this equation over all groups and dividing by $N$ gives

$$w = \beta_o + p \beta_g + p \beta_i.$$  \hfill (A15)

This allows us to rewrite the group effect, the covariance term, in the Price equation above as the between-group variance multiplied by the total derivative of expected group size with respect to the fraction of altruists in a group, namely $\beta_g + \beta_i$:

$$\text{cov}[w_i, p_i] = \sum_j q_j (w_j - w)(p_j - p)$$

$$= \sum_j q_j (w_j - w) p_j$$

$$= \sum_j q_j (p_j - p)(\beta_i + \beta_g) p_j$$

$$= \sum_j q_j (p_j - p) p_j (\beta_i + \beta_g)$$

$$= \sum_j q_j (p_j - p)^2 (\beta_i + \beta_g)$$

$$= \text{var}(p_j)(\beta_i + \beta_g).$$  \hfill (A16)
The second equality follows from the fact that \( \sum_j q_j (w_j - w)p = 0 \) by the definition of \( w \). The third follows by subtracting (A15) from (A14), simplifying, and substituting in the equation. The fourth equation is an algebraic rearrangement. The fifth equation follows from the fact that \( \sum_j q_j (p_j - p) = 0 \), which follows from the definition of \( p \). To evaluate \( E[w_j \Delta p_j] \), we note that

\[
p_j' - p_j = \sum_i q_j' p_{ji} - \sum_i q_j p_{ji} = \sum_i q_j \frac{w_{ji}}{w_j} p_{ji} - \sum_i q_j p_{ji} = \frac{1}{w_j} \sum_i q_j (w_{ji} - w_j) p_{ji},
\]

so

\[
E[w_j \Delta p_j] = \sum_i q_j (w_{ji} - w_j) p_{ji} = \beta_j \sum_i q_j (p_{ji} - p_j) p_{ji} = \beta_j \text{var}(p_{ji}). \tag{A17}
\]

Thus, substituting equations A16 and A17 into the Price equation, we have

\[
w \Delta p = \text{var}(p_j) \beta - \beta_1 \text{var}(p_{ij}). \tag{A18}
\]

Substituting the costs and benefits of the altruistic behavior for the \( \beta \)'s in this equation we have

\[
w \Delta p = \text{var}(p_t)(b - c) - c \text{var}(p_{ji}). \tag{A19}
\]

where, as before, \( w \) is the population-wide average of the number of replicas (which we normalize to unity) and the expectation operator \( E \) indicates a weighted average over groups (the weights being relative group size).

For an example of using the Price equation, suppose a population is composed of two groups that in a given period are of equal size, with the fractions of altruists in each, \( p_1 = 3/4 \) and \( p_2 = 1/4 \), so \( p = 1/2 \). To find the values of \( b \) and \( c \) such that \( p \) will be stationary, we need to equate the average fitness of the two types. Writing \( w_{ij} \) for the fitness of type \( i \) in group \( j \) (\( i = A, N, j = 1, 2 \)) and \( w_i \) for the population average fitness of type \( i \), and ignoring \( \beta_o \), we have

\[
w_A = p_1 w_{A1} + p_2 w_{A2} = (1 - p_1) w_{N1} + (1 - p_2) w_{N2} = w_N. \tag{A20}
\]

1More generally, if \( x \) and \( y \) are random variables and \( y = a_o + a_1 x \) is the regression equation of \( x \) on \( y \), then \( \text{cov}(x, y) = a_1 \text{var}(x) \).
or, using the payoff table in the text, the average fitness of A’s and N’s is the weighted average of their respective fitnesses in the two groups so the condition for \( p \) to be stationary, is given by

\[
\begin{align*}
w_A &= \frac{3}{4} \left( \frac{3}{4} b - c \right) + \frac{1}{4} \left( \frac{1}{4} b - c \right) \\
&= \frac{1}{4} \left( \frac{3}{4} b \right) + \frac{3}{4} \left( \frac{1}{4} b \right) = w_N.
\end{align*}
\]

Solving, we find the values of \( b \) and \( c \) for which \( \Delta p = 0 \), namely, \( c/b = 1/4 \). This means that, given the assumed distribution of A’s and N’s in the two groups, the population frequency of A’s will be stationary if the cost of performing the altruistic act is one fourth the benefit. If we add the further requirement that the size of the total population be constant, so \( w_A = 1 = w_N \) and assuming \( \beta_o = 0 \), we have \( b = 8/3 \) and \( c = 2/3 \).

An equivalent method is simply to use equation (A19), along with the facts that \( \text{var}(p_j) = p_j(1-p_j) = 3/16 \) for \( j = 1, 2 \) and \( \text{var}(p_j) = 1/16 \), so, using (A19), we have from the Price equation 6.2

\[
w \Delta p = (b - c)/16 - 3c/16,
\]

which, for \( w \neq 0 \), gives \( c/b = 1/4 \) as a condition for \( \Delta p = 0 \), reproducing the above result.

Further, reproducing the condition for the stationarity of \( p \), given by (6.9), and using the empirical values from the example gives us:

\[
\frac{c}{b} = F = \frac{\text{var}(p_j)}{\text{var}(p_{ij}) + \text{var}(p_j)} = (1/16)/(3/16 + 1/16) = \frac{1}{4},
\]

as we would expect.

Thus, for values of \( b > 4c \), the frequency of the altruistic trait will grow, exceeding one half in the next period. This occurs because the relative size of the more altruistic group grows, offsetting the decline in the fraction of altruists in each group.

The proliferation of the group-beneficial but individually-costly trait is explained by the group structure of the population, which accounts for the fact that altruists tend to be paired with other altruists more frequently than the population average, despite random pairing within groups. Thus, the probability of meeting an altruist conditional on being an altruist is

\[
P(A|A) = p_1^2 + p_2^2 = 5/8.
\]
This follows from the fact that a fraction \( p_1 \) of all A’s are in group 1 in which the likelihood of being paired with an A is \( p_1 \), and analogously for the fraction \( p_2 \) of A’s in group 2. Non-altruists meet altruists with probability

\[
P(A|N) = (1 - p_1)p_1 + (1 - p_2)p_2 = 3/8.
\]

The difference between these two conditional probabilities, 1/4, is a measure of the degree of positive assortment in this population, and it is the expected advantage enjoyed by the altruistic trait by dint of its favored distribution among groups. Thus, reproducing (6.7) and (6.9), we have an equivalent way of representing (A22):

\[
\frac{c}{b} = P(A|A) - P(A|N) = \frac{1}{4}.
\] (A23)

### A10 Weak Multi-level Selection

We assume all agents are genetically identical and there is clonal reproduction. We assume also that altruist phenotypes (A’s) pass their cultural preference on to the their offspring, as do the non-altruists (N’s). As this is a cultural model, the term fitness means the number of replicas that an individual will make in the next period which may be one, if the individual retains the status quo trait or greater, if the individual is copied by another, or zero if the individual abandons the trait. The fitness of an N due to material payoffs is now \( m_j^N = bp_j \), where \( p_j \) is the fraction of A’s in group \( j \), and the fitness of an A due to material payoffs (we term this “material fitness”) is \( m_j^A = bp_j - c \). Suppose \( w_j^A \) is the average fitness of A’s, including the contribution \( m_j^A \), as well as the net probability that N’s switch to A’s through socialization and from A’s by imitation. Similarly, \( w_j^N \) is the average fitness of N’s, including the contribution \( m_j^N \), as well as the net probability that A’s switch to N’s through imitation. Let \( \gamma_a \) be the rate per non-altruist at which N’s are socialized into A’s, and let \( \gamma_n \) be the rate per unit of payoff difference per altruist at which A’s revert to N’s through imitation. Since the payoff difference between A’s and N’s is \( m_j^N - m_j^A = c \), we have

\[
w_j^A = bp_j + \gamma_a(1 - p_j) - \gamma_n c p_j - c + \beta_o
\] (A24)

\[
w_j^N = bp_j - \gamma_a(1 - p_j) + \gamma_n c p_j + \beta_o
\] (A25)
where $\beta_0$ is baseline fitness. There is then a long-run phenotypic equilibrium when $w_j^A = w_j^N$ with a fraction $p_j^*$ of altruists within the group given by

$$p_j^* = \frac{2\gamma_a - c}{2(\gamma_a + c\gamma_n)}.$$

The average material fitness of a member of group $j$, which depends only on the material payoffs and not the phenotypic movements, is then

$$p_j^* m_j^A + (1 - p_j^*) m_j^N = \frac{(b - c)(2\gamma_a - c)}{2(\gamma_a + c\gamma_n)} + \beta_0.$$

Because in equilibrium this is greater than $\beta_0$ provided $\gamma_a > c/2$, a group with an effective socialization process will have higher mean material fitness than a group without such a process. Therefore any inter-group dynamic that favors wealthier groups will lead to the spread of the socialization process by group selection.

### A11 Cooperation and Punishment with Quorum Sensing

In addition to the specification of parameters in §10.1, we will assume the following.

- **Error rate (e):** We interpret the error rate to be the frequency with which an individual does not cooperate even when punishment for defection is anticipated. This could occur by mistake, or because idiosyncratic circumstances in which the individual’s cost of cooperating is exceptional (e.g., sick children at home that must be cared for) or another reason not included in the model for incurring the cost of being punished. We simulate three values $\epsilon = 0.01$, $\epsilon = 0.1$, and $\epsilon = 0.2$.

- **Group size (n):** The relevant size is the number of individuals of both sexes in a single breeding generation. The average census size of foraging groups in the ethnographic record that face conditions similar to must human groups in the Late Pleistocene (non arctic, non equestrian) is 37 (Marlowe 2005), which implies an $n$ of about 12 to 18, given that the fraction of a foraging population that is of reproductive age may be between a third and a half. To provide a strong test of our model, we simulate the upper end of this range and two other group sizes, $n = 18$, $n = 36$, and $n = 72$. 


• **Number of interactions** ($T$): The appropriate number of interactions depends on the nature of the cooperative activity in question. Opportunities to cooperate in hunting and sharing prey occur thousands of times in a generation, while cooperation in predation or defense may occur but a few times per decade. We assume $T = 10$, $T = 25$, and $T = 50$.

• **Within-group relatedness** ($r$): We simulate two values: $r = 0$ (random group formation) and $r = 0.07$. The value $r = 0.07$ is the average relatedness in a sample of hunter gatherer groups (Bowles 2006). These are varied one parameter at a time from the base case parameter set $b = 2c$; $p = k = q = 1 : 5c$, $a = 2, \epsilon = 0.1$, $n = 18$, $T = 25$, $r = 0$. For the most part we set the base case so that it is the worst case for cooperation, i.e., the benefit cost ratio is only 2, the cost of punishment is the same as the cost of being punished which is 50 individuals to cooperate, and 10% of the cooperators defect by mistake meaning that there are significant costs of punishing even in the long run. Each figure gives the possible equilibria as a function of the $\tau$ value used by Punishers. Solid circles are stable equilibria, and open circles are unstable equilibria marking the boundaries between the basins of attraction of adjacent stable equilibria. These results are surprisingly insensitive to variation in most of the parameter values. For most of the range of parameters simulated, there is a minimum value of $\tau$. For values slightly greater than that value the minimum frequency necessary for Punishers to increase is low, and nearly maximum average fitness is achieved.

For the most part, variation among results can be understood as stemming from a tradeoff between the first-period costs of inducing Nonpunishers to cooperate, and the long run benefit from the cooperation thereby induced. Note that the magnitude of $p$ is determined by $c$. The cost of being punished need only be sufficiently large as to induce Nonpunishers to cooperate. The cost of punishing is scaled relative to $k$. The other parameters fit into two categories. The parameters $a, \epsilon, b, n$ and $T$ affect the cost of punishing and the long run benefit. Their effect can be understood in terms of how they affect this balance. The effect of $a$ is especially strong because it has a very large effect on how the cost of punishing varies as the number of Punishers in the groups increases. The parameter $r$ reflects the population structure, and has large effects when Punishers are rare. Small positive
values of $r$ create a plausible set of conditions that allows punishment to increase when rare.

Figure A9 offers a sensitivity analysis of agent-based simulation described in §10.3. Note that when cooperation is established, the average shirking rate is below 10% even for very high costs of punishing ($k$), and one standard deviation of the shirking rate in distinct runs is about 30% of the average. The same approximate picture obtains for variations in group size and the migration rate. The lower-left panel shows that the $F_{ST}$ for the population depends on group size as expected from standard population genetics. Thus, cooperation is maintained over rather wide parameter ranges, and the model’s reactions to parameter changes occur in the expected direction.
Figure A4. **The cost of punishment** \((k)\) **varies.** Increasing the cost of punishment decreases the cost of inducing Non-punishers to cooperate, and also increases the cost of punishing errors in the longer run. Thus increasing the cost of punishment reduces the range of conditions under which punishment can evolve.
Figure A5. **The error rate** \((e)\) **varies.** Higher error rates increase the frequency and thus the cost of punishment in the long run and thus decrease the benefit of cooperation induced by punishment. As a result, as error rates increase, cooperative equilibria exist only at higher values of \(\tau\).
Figure A6. **Group size (n) varies.** Larger groups increase the cost of punishment of Nonpunishers during the initial period and errors during later periods. Thus increasing $n$ decreases the range of conditions which allow for punishment and cooperation at equilibrium.
Figure A7. **The expected number of periods ($T$) in an interaction varies.** Increases in the expected number of interactions increase the long run benefits of cooperation and thus increase the range of conditions which allow for stable punishment and cooperation.
Figure A8. The relatedness \( r \) among group members vary. In the base case, groups are formed at random, and that means that Punishers are rare, they are alone in groups and can only induce cooperation if they are willing to act alone. Increasing \( r \) means that there is a positive probability that more than one punisher will be together in groups even when Punishers are rare. Thus for small values of \( r \), punishment can increase when rare even when \( r \) is assumed to be between 0.035 and 0.07 the latter being the mean value measured among contemporary foraging populations.
Figure A9. **Sensitivity to Variation of Model Parameters.** Baseline parameters are as in Table 10.1.
From (A26) and the definition of \( \xi_i \) and \( \xi_j \) we calculate that

\[
\frac{da_i^{\text{Nash}}}{d\gamma_i} = \frac{2c\lambda_j}{\lambda_j(4v_1v_2 - c^2\chi^2)^2},
\]

which is positive, so an increase in guilt increases \( i \)'s contribution in the Nash equilibrium. Also,

\[
\frac{da_i^{\text{Nash}}}{dv_i} = \frac{4v_j(2v_j\xi_i + c\chi\xi_j)}{\lambda_j(4v_1v_2 - c^2\chi^2)^2},
\]

which is also positive, so an increase in \( i \)'s shame parameter increases his contribution in the Nash equilibrium. We also have

\[
\frac{da_i^{\text{Nash}}}{d\lambda_j} = \frac{c(\chi\xi_j + 2v_j(1 - \gamma_i - \chi))}{\lambda_j^2(4v_1v_2 - c^2\chi^2)}
\]

so an increase in \( j \)'s level of reciprocity leads to an increase in \( i \)'s contribution in the Nash equilibrium.

We can indeed show that under favorable conditions, and increase in an increased sense of shame may enhance ones own payoffs: i.e., \( d\pi_i^{\text{Nash}}/dv_i > 0 \). For ease of exposition, after taking this derivative, we will set \( v_i = v_j = v \) and \( \lambda_i = \lambda_j = \lambda \) for ease of exposition, and then let \( v = c\chi/2 + \epsilon \) for small \( \epsilon > 0 \). Note that this is possible provided \( \xi_1 \) and \( \xi_2 \) are suitable small, so that \( 0 < \alpha_i^*, \alpha_j^* < 1 \) and the stability conditions remain satisfied. In this case, algebraic manipulation shows that for sufficiently small \( \epsilon \),

\[
\text{sign} \frac{d\pi_i^{\text{Nash}}}{dv_i} = \text{sign} 2c^3\lambda\chi^3(2\lambda - c(2 - \gamma_1 - \gamma_2 - 2\chi))^2,
\]

which is positive.
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Chapter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>3, 4, 8, 9</td>
<td>Group Size</td>
</tr>
<tr>
<td>$b$</td>
<td>3, 8, 9, 10</td>
<td>Benefit</td>
</tr>
<tr>
<td>$c$</td>
<td>3, 8, 9, 10</td>
<td>Cost</td>
</tr>
<tr>
<td>$r$</td>
<td>3, 10</td>
<td>Genetic Relatedness</td>
</tr>
<tr>
<td>$\delta$</td>
<td>3, 10</td>
<td>Continuation Probability</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>3, 4</td>
<td>Error Rate</td>
</tr>
<tr>
<td>$p$</td>
<td>4</td>
<td>Level of Punishment</td>
</tr>
<tr>
<td>$i, j$</td>
<td></td>
<td>Indices for Individuals and/or Groups</td>
</tr>
<tr>
<td>$\delta_{ij}$</td>
<td>6, 8</td>
<td>Probability Individual $i$ in Group $j$ is an Altruist</td>
</tr>
<tr>
<td>$w_{ij}$</td>
<td>6, 8</td>
<td>Fitness of Individual $i$ in Group $j$</td>
</tr>
<tr>
<td>$w$</td>
<td>6, 8</td>
<td>Average Fitness</td>
</tr>
<tr>
<td>$w_j$</td>
<td>6, 8</td>
<td>Average Group $j$ Fitness</td>
</tr>
<tr>
<td>$p_j$</td>
<td>6, 8</td>
<td>Fraction of Group $j$ who are Altruists</td>
</tr>
<tr>
<td>$\beta_o$</td>
<td>6, 8</td>
<td>Baseline Replication Rate</td>
</tr>
<tr>
<td>$\beta_g$</td>
<td>6, 8</td>
<td>Effect of $p_j$ on $w_{ij}$</td>
</tr>
<tr>
<td>$\beta_i$</td>
<td>6, 8</td>
<td>Effect of $p_{ij}$ on $w_{ij}$</td>
</tr>
<tr>
<td>$\beta_G$</td>
<td>6, 8</td>
<td>$\beta_i + \beta_g$</td>
</tr>
<tr>
<td>$F$</td>
<td>6</td>
<td>Wright’s Variance Ratio</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>8</td>
<td>Conflict Probability</td>
</tr>
<tr>
<td>$\lambda, \lambda_A$</td>
<td>8</td>
<td>Probability of Deme Survival</td>
</tr>
<tr>
<td>$\mu$</td>
<td>9</td>
<td>Mutation Rate</td>
</tr>
<tr>
<td>$m$</td>
<td>6, 7, 8, 10</td>
<td>Migration Rate</td>
</tr>
<tr>
<td>$\tau$</td>
<td>8</td>
<td>Reproductive Leveling</td>
</tr>
<tr>
<td>$\tau$</td>
<td>10</td>
<td>Quorum Level</td>
</tr>
<tr>
<td>$\xi$</td>
<td>8</td>
<td>Segmentation Rate</td>
</tr>
<tr>
<td>$a$</td>
<td>10</td>
<td>Returns to Scale in Punishing</td>
</tr>
<tr>
<td>$k$</td>
<td>10</td>
<td>Cost of Punishing</td>
</tr>
<tr>
<td>$n_p$</td>
<td>10</td>
<td>Number of Punishers</td>
</tr>
<tr>
<td>$p$</td>
<td>10</td>
<td>Cost of being Punished</td>
</tr>
<tr>
<td>$q$</td>
<td>10</td>
<td>Cost of Signaling</td>
</tr>
<tr>
<td>$T$</td>
<td>10</td>
<td>Number of Periods of Interaction</td>
</tr>
<tr>
<td>$\rho$</td>
<td>10</td>
<td>Reproduction Rate</td>
</tr>
<tr>
<td>$W_0$</td>
<td>10</td>
<td>Baseline Fitness</td>
</tr>
<tr>
<td>Symbol</td>
<td>Chapter</td>
<td>Meaning</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>( \eta )</td>
<td>11</td>
<td>Imitation Rate</td>
</tr>
<tr>
<td>( f )</td>
<td>11</td>
<td>Fitness Gain from C phenotype</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>11</td>
<td>Rate of Oblique Transmission</td>
</tr>
<tr>
<td>( s )</td>
<td>11</td>
<td>Fitness cost of A phenotype</td>
</tr>
<tr>
<td>( u )</td>
<td>11</td>
<td>Fitness Cost of ( a ) allele</td>
</tr>
<tr>
<td>( \mu_{ij} )</td>
<td>12</td>
<td>Punishment of ( j ) by ( i )</td>
</tr>
<tr>
<td>( a_i )</td>
<td>12</td>
<td>( i ) ’a Contribution</td>
</tr>
<tr>
<td>( \chi )</td>
<td>12</td>
<td>Productivity of Contributing</td>
</tr>
<tr>
<td>( \gamma_i )</td>
<td>12</td>
<td>Strength of ( i )’s Moral Standard</td>
</tr>
<tr>
<td>( \lambda_i )</td>
<td>12</td>
<td>Material Payoff to ( i )</td>
</tr>
<tr>
<td>( \nu_i )</td>
<td>12</td>
<td>Strength of ( i )’s Shame</td>
</tr>
<tr>
<td>( \pi_i )</td>
<td>12</td>
<td>Material Payoff to ( i )</td>
</tr>
</tbody>
</table>
REFERENCES


References


Benedict, Ruth, Patterns of Culture (Boston: Houghton Mifflin, 1934).


References


References


References


References


304 References


Cybulski, Jerome, “Culture Change, Demographic History and Health and Disease on the Northwest Coast,” in Clark Spencer Larsen and George Milner (eds.) In the Wake of Contact: Biological Responses to Conquest (Wiley-Liss, 1994) pp. 75–85.

Daly, Martin and Margo Wilson, Homicide (New York: Aldine de Gruyter, 1988).


___. *The Extended Phenotype: The Gene as the Unit of Selection* (Oxford: Freeman, 1982).


Fehr, Ernst and Andreas Leibbrandt, “Cooperativeness and Impatience in the Tragedy of the Commons,” 2010. University of Zurich.


References


Fix, Alan, Migration and Colonization in Human Microevolution (Cambridge: Cambridge University Press, 1999).


310 References


Gamez, Stefany Moreno, Jon Wilkins, and Samuel Bowles, “Cosmopolitan Ancestors: Simulations Calibrated with Genetic and Ethnographic Data Show that Prehistoric Populations were not Small and Isolated,” 2010. Santa Fe Institute.


Gilens, Martin, Why Americans Hate Welfare (University of Chicago Press, 1999).


References


References


Hiatt, L. R., Kinship and Conflict: A Study of an Aboriginal Community in Northern Arnhem Land (Canberra: Australian National University, 1965).


Hrdy, Sarah Blaffer, Mother Nature: Maternal Instincts and How They Shape the Human Species (New York: Ballantine, 2000).


References


318 References


References


Mead, George Herbert, Mind, Self, and Society (Chicago: University of Chicago Press, 1967[1934]).


Morgan, J., The Life and Adventures of William Buckley: Thirty-two Years a Wanderer Amongst the Aborigines (Canberra: Australia National University Press, 1979[1852]).

322 References


References


—, *The Theory of Moral Sentiments* (New York: Prometheus, 2000 [1759]).


References


Tesaroff, Leigh and Kenneth L. Judd, Handbook of Computational Economics II: Agent-Based Computational Economics (North-Holland, 2006).


330 References


332 References

Index

aboriginal cultural practices, 127
Adam Smith, 9, 256
agent-based model, 39, 112, 163, 179, 192, 201, 232, 258–267
Alexander Field, 11
Alexis de Tocqueville, 12
Alfred, Lord Tennyson, 11, 173
altruism, 13
as signal of quality, 48–49
definition, 30, 277–279
evolution of, 176
kin, 30–33, 253
reciprocal, 5, 33–48, 57, 117, 253
warm glow, 82
altruistic cooperation, 5
altruistic punishment, 81
Amos Zahavi, 48
Antonio Damasio, 240
assortative interaction, 199
assortment, 31
positive, 31, 116–119
assurance game, 18
Astrid Hopfensitz, 165
Avner Shaked, 100
Barry Hewlett, 217
behavior
and cultural difference, 92
context-dependent, 74
framing effects, 102
insider-outsider, 95
institutional cues, 94
parochial, 97
variability, 216
beliefs, 13–15
beliefs, preferences, and constraints, 13
beliefs, preferences, and constraints model, 13
Benedikt Herrmann, 83
best response, 17, 58
Bruce Winterhalder, 254
cataclysmic demographic events, 128
character virtues, 16
Charles Darwin, 173
Charles Horton Cooley, 16, 104
Christian Thöni, 83
Christopher Boehm, 11, 25, 126, 143, 151, 248
coevolution of parochialism and altruism, 179
common pool resource game, 16, 95
complete contracts, 52
conformist cultural transmission, 107, 114, 149, 210, 218
constraints, 13–15
continuation probability, 61, 62, 267–268
contrite behavior, 41
convention, 149
cooperation, 4
unraveling of, 68
within-group, 7
Index

cooperative equilibrium, 66
correlated equilibrium, 64
correlating device, 65
cultural transmission, 21, 214
    biased, 226
    oblique, 218
    vertical, 220
culture, 19
David Sloan Wilson, 11
deme, 108, 132, 152
Devesh Rustagi, 74
dictator game, 73, 74, 91
docility, 218
dominant strategy, 17
dominant strategy equilibrium, 17
dynamical systems, 274–277
dynamically irrelevant equilibrium, 52, 62

Elizabeth Phelps, 97
Elke Renner, 85
Elliott Sober, 11
Eors Szathmary, 151
equilibrium
    correlated, 64
    choreographer, 66
dominant strategy, 17
dynamically irrelevant, 62
Nash, 18, 52, 64, 263–267
phenogenotypic, 228
phenotypic, 285
selection, 24
stability of, 229
Eric Alden Smith, 48, 254
Erika Seki, 73
evolution of cooperation
    hunting, 251
    war, 173
evolutionarily stable strategy, 34, 266

evolutionary game, 265, 267
replicator dynamic, 265
evolutionary universals, 148
execution error, 40
experimental labor market, 77

F. Y. Edgeworth, 14
Folk Theorem, 52
Frank Marlowe, 126
Frans de Waal, 11
free-riding, 80

G. R. Price, 28, 29
game, 263
    best response, 264
dictator, 74
dominant strategy, 17, 264
evolutionary, 265
    extensive form, 264
gift exchange, 76
    mixed strategy, 263
    Nash equilibrium, 54
    non-cooperative, 17
    one-shot, 54
    public goods, 39, 46, 81
    pure strategy, 263
    repeated, 54, 211
    stage, 265
    strategic form, 263
    third party punishment, 89
    ultimatum, 70
Garrett Hardin, 10
gene-culture coevolution, 20, 150, 254
genetic drift, 131, 132, 158, 226
George Herbert Mead, 104
George Price, 109
gift exchange game, 37, 76
globally stable, 221
group competition, 106
group conflict, 7, 31, 108, 125, 130, 141, 148, 170, 173, 181, 189, 253
group selection, 26, 108, 150, 151
cultural, 148
    strong, 108, 121
    weak, 108, 232
guilt, 240
    H. L. Mencken, 3
    Hamilton’s rule, 30, 35, 116
    helping, 12
    Henry Tajfel, 95
    Herbert Simon, 218
    Hobbes, Thomas, 10
    human conscience, 248
    human cooperation
        evolution of, 121
        origins, 251
    image scoring, 268–269
    in-group favoritism, 175
    inbreeding coefficient, 111
    inclusive fitness, 28, 30
    indirect reciprocity, 49
    individual learning, 21
    information
        imperfect, 53
        perfect, 53
        private, 53
        public, 53
    insider favoritism, 173
    interaction
        dyadic, 38
        in large groups, 39
    intergroup conflict, 171
    internalization of norms, 24, 104, 214–238, 245, 246
    invisible hand, 9
    Irenaus Eibl-eibesfeldt, 151
    J. B. S. Haldane, 107
    James Friedman, 33
    James Habyarimana, 97
    Jean-Jacques Rousseau, 124
    Jeffrey Carpenter, 73
    Jeremy Bentham, 217
    Jessica Stern, 14
    John Harsanyi, 63
    John Maynard Smith, 29, 31, 151
    John Nash, 63
    John Stuart Mill, 236
    John Tooby, 124
    Jon Wilkins, 131
    Joshua Epstein, 260
    Julian Huxley, 21
    Jung Kyoo Choi, 165
    Karl Pearson, 173
    Karl Popper, 21
    Ken Binmore, 100
    Kenneth Arrow, 256
    kin, 31
    kin altruism, 31
    kin selection, 31
    laboratory experiment, 71
        external validity, 71
    Lanchester’s Law, 193
    learning
        individual, 21
        social, 21
    Leda Cosmides, 124
    legitimacy, 83
    Linnda Caporeal, 218
    looking-glass self, 104
    Luigi Luca Cavalli-Sforza, 20, 29
    Maarten Voors, 188
    Mancur Olson, 10
    Marcus Feldman, 20, 29
    market-like experiments, 256
Index

Martin Sefton, 85
mating, 125
mating practices, 130, 131
monogamous, 7
mating system, 133
Matt Ridley, 11
Michael Spence, 48
migration, 66, 106, 201, 233, 235
between-group, 131
effective rate, 132
long and short, 128
long distance, 141
non-random, 133
random, 133
rate, 132
mixed strategy, 62, 63, 196, 229
model
agent-based, 112, 179, 258–267
self-interest-based, 211
model standing
alternatives to, 268–269
monotone dynamic, 64
moral outrage, 8
moralistic aggression, 99, 249
strong, 113
weak, 113
mutation, 66, 233
mutualism, 5, 36, 124
myopia, 246
Nash best response, 63
Nash equilibrium, 18, 52, 63, 64, 100, 101, 186, 244, 263–267
dynamically irrelevant, 52
mixed strategy, 63
natural selection, 5, 15, 28, 105, 198, 216
niche construction, 22
non-cooperative game, 17
norms, 217
internalization, 7
internalization of, 214–238
internalized, 217
Norse peoples, 175
oblique conformist transmission, 218
one-shot, 174, 242
osteological evidence, 138
other-regarding, 95
other-regarding behavior, 79, 91, 104
other-regarding individuals, 90
other-regarding motives, 92, 212
other-regarding preferences, 15, 17, 48, 68, 71, 92, 102, 125, 255, 277
parochial altruism, 97, 173–189
parochial networks, 185
parochialism, 175
Parthik Panchanathan, 46
partner treatment, 37, 81
payoff monotonic dynamic, 23
payoff-based updating, 230
perceptual errors, 53
perfect stranger treatment, 81
Peter Hammerstein, 36
Peter Kropotkin, 11
Peter Richerson, 20
phenogenotypes, 223
phenotypic expression, 23, 216
Pleistocene, 25
Pleistocene climate records, 125
Polly Wiessner, 27
positive assortment, 31, 112, 116–119, 154, 157, 175, 181, 203, 208
preferences, 13–15, 217
complete, 13
consistent, 13
internalized, 217
Index 337

other-regarding, 15, 16
self-regarding, 15, 16
social, 6, 16
  altruistic, 7
transitive, 13, 91
Price equation, 279–284
prisoner’s dilemma, 9, 16, 17, 18, 19, 33, 54, 62, 65, 76, 79, 95, 99, 100, 185, 186, 264, 265, 272
public goods game, 16, 39, 40, 46, 52, 72, 73, 80, 81, 84, 88, 93, 98, 118, 185, 186, 240, 242
  experimental, 78
  multi-round, 79
  with punishment, 39, 83, 88, 100
public signals, 41
punishment
  absence of second order, 144
  altruistic, 81, 191
  coordinated, 144, 192
  counter-productive, 191
  ostracism, 7, 142, 144, 145, 149, 249
  peer, 146
  quorum-based, 208
  shunning, 7
  symbolic, 88
  third-party, 89
quorum sensing, 193–210, 285–294
rational, 63, 65, 265
rationality, 14, 90, 91
  of social preferences, 91
recent common descent, 69
reciprocity
  indirect, 90
replicator dynamic, 64, 265, 266, 267
replicator equation, 220
resource sharing
institutionalized, 171
Richard Alexander, 12, 44
Richard Boyd, 46
Richard Dawkins, 11, 12, 21, 28, 123, 124
Richard Klein, 251
Richard Lee, 142
Richard Wright, 11
Robert Axelrod, 29, 34
Robert Axtell, 260
Robert Boyd, 20, 252
Robert Edgerton, 215
Robert Sugden, 44
Robert Trivers, 28, 33, 123, 216
sanctioning selection, 213
Sarah Hrdy, 11
selection
  among equilibria, 24
  between-deme, 153
  between-group, 163, 191
  cultural, 175
  genetic, 22, 175
  group, 108
    cultural, 148
  natural, 22
  sanctioning, 212
  within-group, 106, 161, 165, 215
selective emigration, 154, 176
self-regarding preferences, 15, 17, 95
shame, 8, 239
  susceptibility to, 248
sick societies, 216
signal
  imperfect, 53
  imperfect public, 56
  perfect, 53
  private, 53, 60
  public, 53
  signal errors, 53
Index

signaling
  costly, 48
Simon Gächter, 83
social, 38
social dilemma, 16, 53, 73, 78, 85, 98, 186
social emotions, 8, 239–249
evolution of, 245
social learning, 21
social norm, 64, 65
social norms, 3, 26, 66, 67, 86, 89, 90, 123, 142, 143, 190, 231, 232
  internalization of, 217
social order
  foundations, 142
social preferences, 15, 70
socialization, 7, 214–238
stage game, 52, 54, 55, 260, 265
standing model, 44, 268–269
standing strategy, 44, 46, 47, 268
state of nature
  Hobbesian, 147
stationary distribution, 40
Stefany Moreno Gamez, 131
Steven Frank, 151
stranger treatment, 37, 81
strategy
  dominant, 17
  mixed, 62
  uninvadable, 34
strong reciprocity, 26, 76–78, 123, 190, 210, 253

Talcott Parsons, 148
Thomas Hobbes, 10, 15, 147
Thomas Schelling, 259
Thorstein Veblen, 15
time discount factor, 267
tipping groups, 206
tit-for-tat, 124, 209

nice, 33
Toshio Yamagishi, 96
tragedy of the commons, 9
transitive preferences, 91
transmission
  cultural, 114
  gene-culture, 223
trigger strategy, 54, 55, 271
Truman Bewley, 14
trust game, 72, 96, 174
ultimatum game, 70, 73, 93, 94, 98, 100, 239
uninvadable, 34

variance
  between-deme, 110
  within-deme, 110
variance ratio, 111
vertical transmission, 220
volatility
  climate, 141
  temperature, 128

war
  archeological evidence, 137
  mortality, 140
  pre-modern, 139
  prehistoric, 136
  warfare, 7
warm glow altruism, 83
welfare economics
  Fundamental Theorem of, 51
William Durham, 20
William Hamilton, 13, 28, 34, 106