

3

Social Preferences

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

3.1 Introduction

Altruistic cooperation is common because people are motivated by *social preferences*; that is they care about the well-being of others, and value fairness and other norms of decent behavior. Because our explanation holds that social preferences are the proximate cause of altruistic cooperation, the task of this chapter is to establish that these preferences are indeed common. This fact about the proximate causes of cooperation poses the challenge of evolutionary explanation to which the subsequent chapters are addressed: in light of what we know about other animals and about human evolution, how could social preferences have become common among humans?

[Sam] Our conception of individual choice may be termed the “beliefs, preferences, and constraints” model. We use the term “preferences” in a way that is standard among decision theorists. Given a set of feasible actions, what people do in any situation therefore depends on their preferences and their beliefs. We use the term *constraints* for the set of feasible actions an individual may take in a given situation. [Sam] *Beliefs* are an individual's understandings of the nature and causal structure of the world in which he lives, including the relationship between actions and outcomes. *Preferences* are reasons for behavior, that is attributes of individuals—other than beliefs and capacities—that account for the actions they take in a given situation. Preferences may be described as an ordering of states (technically, a preference function) on two conditions: that preferences are *complete* (all states can be ordered) and preferences are *transitive*; that is, *consistent*, so that if one prefers A to B and B to C one does not also prefer C to A. Prefer-

ences are the results of a heterogeneous mélange 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 one's affective relationships with others. To say that a person acts on his preferences means only that knowledge of the preferences would be helpful in providing a convincing account of the actions (though not necessarily the account which would be given by the actor, for as is well known individuals are sometimes unable or unwilling to provide such an account).

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 seek to 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 are thought to capture important influences on individual behavior.

A version of this model, incorporating the behavioral assumptions sometimes summarized as *Homo economicus*, is rapidly becoming standard throughout the human behavioral sciences. F. Y. Edgeworth, a founder of the neoclassical paradigm in 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 (one could have transitive and complete altruistic or masochistic preferences) but it is commonly treated as axiomatic in economics (and sometimes 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 individual self-interest coupled with the belief that others are also motivated by self interest.

But predictions based on Edgeworth's self-interest axiom are often falsified. 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) study of why firms do not cut wages during recessions, Jessica Stern's (2003) study of terrorist violence and the Hagoromo Society's study of kamikaze pilots (Hagoromo Society 1973). The behavioral experiments that we will explain presently provide additional evidence against Edgeworth's view.

While standard practice in most sciences, our reliance on experimental findings is at variance with the economist's usual motivation for the self-interest axiom, namely that it is self-evident, with the fallback assertion being that evolution 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 fallback assertion is false. Everyday observation of others as well as introspection suggests that social preferences are important. Compelling evidence of social preferences comes from real world behaviors that are inexplicable in terms of self-interest without resort to extensive *ad hoc* reasoning.

Examples include the importance of fairness motives in wage setting and other exchanges (Bewley 2000, 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 (Fong 2001, Gilens 1999, Fong, Bowles and Gintis 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, Erand and Feinstein 1998), participating in various forms of collective action with little expectation of personal benefit (Wood 2003), and conforming to norms and laws in cases where one's transgression would be personally advantageous and would not be detected.

But, the experiments add critical information for, 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. Of course, experimental results in the laboratory would

not be very interesting if they did not reflect real-life behavior. While much more work in this area remains to be done, there are strong and consistent indications that the external validity of experimental results is high. For instance, Binswanger (1980) and Binswanger and Sillers (1983) used survey questions concerning attitudes towards risk and experimental lotteries with real financial rewards to predict successfully the investment decisions of farmers. Glaeser, Laibson, Scheinkman and Soutter (2000) explored whether experimental subjects who trusted others in what is called the trust game 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. Genesove and Mayer (2001) showed that loss aversion determines seller behavior in the 1990's Boston housing market. Condominium owners subject to nominal losses set selling prices equal to the market rate plus 25% to 35% of the difference between their purchase price and the market price, and they sold at prices 3% to 18% of this difference. These findings show that loss aversion is not confined to the laboratory alone, but affects behavior in a market on which very high financial gains and losses can be made.

Similarly, Karlan (2005) used trust and public goods games to predict the probability that loans by a Peruvian microfinance lender would be repaid. He found that individuals who were "trustworthy" in the trust game were less likely to default. Also, Ashraf, Karlan and Yin (2007) 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 and only those who exhibited loss aversion in a laboratory survey also exhibited loss aversion when faced with real-life wage changes.

We begin by defining social preferences and relating them to the common terms: altruism, mutualism, and self-interest. We then review several major results of literally hundreds of behavioral experiments around the world over the past two decades. We conclude with the evolutionary puzzle posed by these data.

3.2 Social Preferences and Altruistic Cooperation

Preferences are based on the evaluations of the consequences of actions. In contrast to the standard economic assumption of self-regarding preferences based on evaluations of states that refer to oneself, we stress *other-regarding preferences*; that is, 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 Hobbes called the desire for “eminence” or Thorstein Veblen’s “pecuniary emulation” exemplified by a desire to “keep up with the Joneses” (Veblen [1899]/1934). A key aspect of other-regarding preferences is that one’s evaluation of a state depends in part on how it is experienced by others. Also important are ethical preferences, and these need not be other-regarding. One could be honest because one is other-regarding and seeks to avoid imposing costs on others by deceiving them. But honest behavior could be entirely self-regarding, the cost being endured in this case 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 one’s ethical duty. But social preference, generosity, fair-mindedness, and a commitment to honesty, for example, often motivate people 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 helping 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 by the effects of one’s actions on others.

To clarify the distinction between self-regarding and other-regarding preferences, consider perhaps the most famous of all experimental games (§A2), the *prisoners’ dilemma*, with payoffs (the row player’s first) shown in Fig-

ure 3.1. 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 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).

	C	D
C	10,10	0,15
D	15,0	5,5

Figure 3.1. The Prisoners' Dilemma

A *self-regarding* player cares only about the payoff to himself. He thus reasons as follows. "If my partner cooperates, I get \$15 by defecting 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-interested person, Defect is a *dominant strategy*, that is, the best response 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 equilibrium*.

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 equilibria in this transformed game. Social preferences convert a prisoners' dilemma material payoff structure

into an *assurance game*. Which of the two equilibria will obtain depends on the players' beliefs about what the other will do.

In one shot prisoners' dilemma experiments, the rate of cooperation is commonly between 40% and 60% (Fehr and Fischbacher 2002). Many subjects prefer the mutual cooperation outcome over the higher material payoff they would get by defecting on a cooperator. Moreover, as we will see, many defect not because they are self-regarding, but to avoid risking being exploited by a selfish partner. [Sam] The idea that many people prefer to cooperate in a prisoner's dilemma provided their partner cooperates as well may be captured by supposing an altruistic cooperator enjoys a psychic benefit equal to \$6 if both he and his partner cooperate. The transformed game is then depicted in Figure 3.2, where $x = 10$ if player 2 is self-regarding, and $x = 16$ if player 2 also prefers to cooperate. Note that if player 2 is self-regarding, he still has a dominant strategy of defecting, in which case player 1, although preferring the cooperative solution, will maximize the sum of his material and psychic payoffs by defecting. On the other hand, if player 2 also prefers cooperating, so $x = 16$, there are now two Nash equilibria—the mutual defect DD and the mutual cooperate CC. If each player believes the other is a cooperator, or if either one believes the other is a cooperator and is allowed to move first, then CC will be the resulting equilibrium.

	C	D
C	16, x	0,15
D	15,0	5,5

Figure 3.2. The Prisoners' Dilemma with Cooperative Preferences

Consider another experimental game, the so-called *ultimatum game* (Güth, Schmittberger and Schwarze 1982). In this one-shot, anonymous game, 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 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. 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 the minimum possible positive amount, and this will be accepted.

However, when actually played, the self-regarding outcome is almost never attained and rarely is approximated. As many replications of this experiment in over thirty countries have documented, under varying conditions and with varying amounts of money, proposers routinely offer respondents very substantial amounts, 50% of the total generally being the modal offer. Respondents frequently reject offers below 25% (Roth, Prasnikar, Okuno-Fujiwara and Zamir 1991, Camerer and Thaler 1995). The fact that positive offers are commonly rejected shows that respondents have other-regarding preferences, and the fact that most proposers offer \$4 or \$5 shows that either proposers have other-regarding preferences, or they at least believe respondents have other-regarding preferences. 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 if it means giving up some money to do so.

Are there other plausible interpretations of this behavior? One might suggest that subjects simply did not understand the game. This is not very plausible, because the game is 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 should eventually learn to accept any positive offer. In fact, generally the rejection rate does not decline with repetition.

Another possibility is that other-regarding preferences emerge when the stakes are low, but would disappear with higher-stakes games. This is not the case. Several researchers have tested this proposition, and found it not to hold (Roth et al. 1991, Hoffman, McCabe and Smith 1994a, Straub and Murnighan 1995, Cameron 1999). Slonim and Roth (1998), however, show that if the ultimatum game is repeated ten times with different partners each time, there is a small but significant tendency for rejections to decline when the stakes are very high (about ten days' wages in Slovakia), but not otherwise. The fact that in this experiment rejections vary with the stakes of the game does not indicate the lack of social preferences, but only that subjects take account of the costs of their behaviors.

A third possibility is that if the anonymous, non-repeated interactions characteristic of experimental games were not a part of everyday life or our evolutionary history, we would not expect subjects in experimental games

to behave in a self-regarding manner in these unique situations. Rather, we might expect subjects to confuse the experimental environment with a repeated interaction, and 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 wisdom that their experiences and their ancestors have conveyed. We will see powerful evidence of this below. But, among these is not the inability to tell a one-shot from a repeated interaction or rule that says “act the same way in both.” 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, 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). Evidence of this is clear in Figure 4.2.

Other data support the notion that responders reject positive offers, not for self-interested reasons, 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 never reject offers, and proposers make considerably smaller (but still positive) offers. When asked, subjects who have rejected low offers often express anger at the proposer and a desire to penalize unfair behavior.

Thus, the concern that rejections of positive offers by ultimatum game respondents are attributable to confusion, low stakes, or subjects’ mistakenly framing the interaction as a repeated one appear to be misplaced. Rejection of positive offers is evidence of other-regarding or moral preferences.

We define a *social dilemma* as an interaction involving several people in which the utility maximizing behavior of a self-regarding person results in an outcome that is socially inefficient in the (Pareto) sense that there exists some other feasible allocation such that at least one member could be better off while no member would be worse off. Examples of social dilemmas are the prisoners’ dilemma, the public goods game (§4.4), sometimes termed an *n*-person prisoners’ dilemma, the so-called “war of attrition” and other arms race interactions, the tragedy of the commons and the common pool resource problem 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* on the cooperation of

other group members if he benefits from the actions of others, while himself behaving as a self-regarding individual. The puzzle of cooperation is to understand why altruistic cooperation is so common in social dilemmas, and why free riding does not always lead to the unravelling of cooperative behavior.

We use the term *strong reciprocity* to describe the motives for much of the cooperation observed both in experimental and natural settings. 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. We turn now to experimental evidence for strong reciprocity and other kinds of social preferences.

3.3 Strong Reciprocity is Common

In dyadic interactions, strong reciprocity is common. 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 ultimatum and prisoners' dilemma games. In addition, experiments by Hayashi, Ostrom, Walker and Yamagishi (1999) and Watabe, Terai, Hayashi and Yamagishi (1996) investigated a "sequential" prisoners' dilemma where subjects were told their partner's choice and then decided whether to cooperate or free-ride. When told their partner did not cooperate, 87% of Japanese subjects and 100% of American subjects also did not cooperate. When informed their partner cooperated, 75% of the Japanese subjects and 61% of the American subjects cooperated. Similar results are reported by Morris, Sim and Giroto (1998), Kiyonari, Tanida and Yamagishi (2000) and McCabe, Smith and LePore (2000).

Another experiment suggesting that strong reciprocity is common is the "experimental labor market" investigated by Fehr, Gächter and Kirchsteiger (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., $c', c'' > 0$). 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 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 wage is offered them. 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 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 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 employers, 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. Finally, “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, Königstein and Kessler 2004).

3.4 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. The experimenter then 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. 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, in which case each subject then earns just \$10. Because the experimental public goods game is a version of the n -person public goods game (§4.4), it has the property that the dominant strategy for each player is to defect.

In fact, 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 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 and non-repeated 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 fail to contribute on the very first round. Andreoni (1988) tested this prediction, and found it to be inaccurate. 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 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 can be found an experiment by Page, Putterman and Unel (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 period, for an average contribution rate of 38% of the endowment over the twenty periods. 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 relatively inoperative.

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.

3.5 Atruistic 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), and Yamagishi (1988a), (1988b), (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. They ensured that group composition changed in every period so subjects knew that costly retaliation against low contributors could not possibly confer any pecuniary benefit to those who punish.

Fehr and Gächter (2000) used six- and ten-round public goods games 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. 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 3.3. 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 intriguing because in the stranger and perfect stranger treatment, punishing low contributors is no different from contributing to the

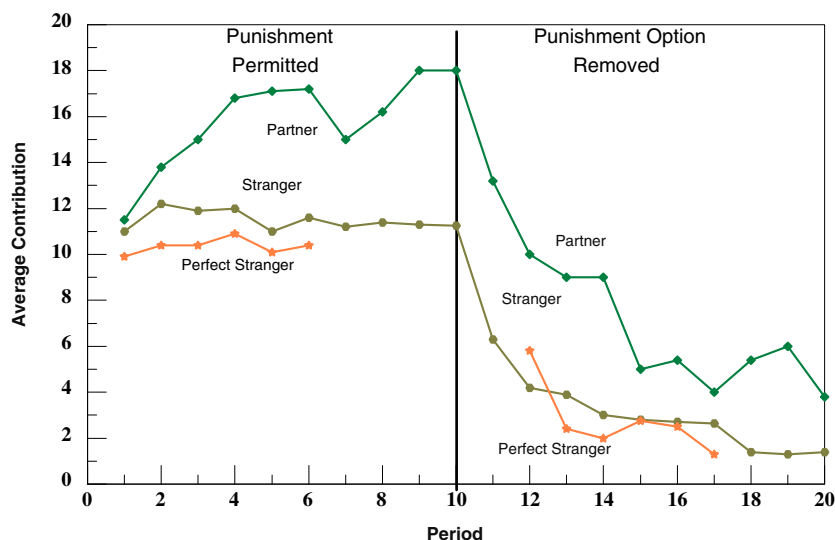


Figure 3.3. Average Contributions over Time in the Partner, Stranger, and Perfect Stranger Treatments when the Punishment Condition is Played First (Fehr and Gächter, 2000). Results are similar when the punishment condition is played second.

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 while once punishment is permitted they avidly engage in the altruistic activity of punishing low contributors.

3.6 People Respond to Symbolic Punishment

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, Noussair, Tucker and Villeval (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, Page and Putterman (2006) compared the usual baseline public goods game with a “chat-room” situation in which group members (group size was four) 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 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.

Recently, Rege and Telle (2001) provided additional experimental evidence suggesting that social rewards and punishments, even in the absence of material consequences, induce cooperative behavior. The experimenters showed this in the context of a ten-person public goods experiment in which every dollar contributed to the public good increases the material payoff of each of the ten group members by 20 cents; i.e. the contributor loses 80 cents. In the baseline condition of this experiment subjects’ contribution to the public good remains anonymous, while in the “approval” condition, both the other subjects and the experimenter can observe each subject’s contribution. Although the subjects were strangers to each other and to the experimenters, in the baseline condition subjects contributed 34% of their endowment to the public good, while in the approval condition the contributions were twice as high.

A plausible interpretation of this is that in the approval condition subjects’ self-image is tarnished by the disapproval of the other group members. This interpretation is supported by Gächter and Fehr (1999), who 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.

3.7 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 (2004) studied “indirect reciprocity” and “strategic reputation-building” in an experimental helping game (§4.5). 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. Self-regarding types will mimic the behavior of strong reciprocators to receive favorable treatment from them.

3.8 Other-regarding 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. These preferences imply observable trade-offs, depending on the cost of behaving morally. Experiments confirm that the higher the cost of moral behavior, the less its frequency, all else taken as given.

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 rational action.

Andreoni and Miller (2002) tested the “rationality” of moral choices by asking 176 subjects to play a version of the so-called “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 π_s , while transferring an amount π_o to B, such that the budget equation $\pi_s + p\pi_o = m$. 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, Horowitz, Savin and Sefton 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, and these violations were almost all very mild. 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 (2000), Anderson and Putterman (2004) 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. 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.

3.9 Institutions and Cultures Matter

The extent to which strong reciprocity and other social preferences occur depends on the institutions framing social interactions and shaping the process of social learning. Independently of the cost effects referred to above, people who behave as strong reciprocators in one situation may behave in entirely self-regarding ways in another. Strong reciprocity is one behavioral capacity in an individual's repertoire of behaviors, not a predisposition dictating a particular kind of behavior in all situations. Partly because of their institutional differences, cultures differ in the extent to which strong reciprocity is a common behavior.

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, Boyd, Bowles, Camerer, Fehr, Gintis and McElreath 2001, Henrich, Boyd, Bowles, Camerer, Fehr and Gintis 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, while even splits were commonly accepted, and 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 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 average proposal was 58 percent 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). Moreover, 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.

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, McCabe, Shachat and Smith 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 games payoffs. Yet the same willingness to adopt a dominated strategy in the interests of the public good is uncommon in the standard public goods game, without punishment. Our interpretation is that the game structure, like institutions in natural settings, conveys information about appropriate behavior and influences beliefs about the actions of others.

Similarly, in the ultimatum game, people in the role of proposer 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, 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. In the role of proposer, considerations of fair treatment are apparently not salient, while in the role of responder, they are.

3.10 Behavior is Conditioned on Group Membership

In experimental and natural settings, people often behave differently towards others, depending on the linguistic, ethnic, racial, 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 race and ethnic identification, and religion (Berscheid and Walster 1969, Cohen 1977, Kandel 1978, Tajfel, Billig, Bundy and Flament 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 (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 prisoners' dilemma and common pool resource 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 recent experiments by 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 expect reciprocation from in-groupers and not from out-groupers. In contrast to the "minimal group" experiments favored by psychologists and sociologists, behavioral economists have the so-called trust game introduced by Berg, Dickhaut and McCabe (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.

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 racial, religious, or linguistic identity of their partner. Fershtman, Gneezy and Verboven (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 from ubiquitous, however. In another study using Belgian subjects, Bouckaert and Dhaene (2003) found no evidence of either type of discrimination in a trust game played by small businessmen of Belgian and Turkish origin. Other studies suggest that in some circumstances, ingroup favoritism is quite limited or even absent.

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, O'Connor, Cunningham, Funayama, Gatenby, Gore and Banaji (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.*). 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 8.

3.11 Conclusion

The most parsimonious and compelling explanation of behavior in the ultimatum, public goods, 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, Rilling, Aronson, Nystrom and Cohen 2003). Camerer, Loewenstein and Prelec (2005) comment: “It is irresistible to speculate that the insula is a neural locus of the distaste for inequality and unfair treatment...”

Our inference that subjects enjoy cooperation is based on a series of experiments in which mutual cooperation outcomes are associated with elevated activity in one of the reward-related areas of the brain, the striatum. Rilling, Sanfey, Aronson, Nystrom and Cohen. (2004) found that mutual cooperation along with a monetary payoff enhances striatum activity 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, Fischbacher, Treyer, Schellhammer, Schnyder, Buck and Fehr (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, an area of the brain involved in processing rewards resulting from a decision. 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 prisoners 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. Our 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 strongly supports Trivers' 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.

The puzzle of human cooperation is this: why are social preferences sufficiently common to sustain the remarkable levels of cooperation observed among humans, even in large groups in which people are substantially unrelated genetically, and in settings in which self-regarding preferences do not support cooperation?

The evolutionary origins and persistence of social preferences among humans is of interest for three reasons. First, asking how we could have become the cooperative species that we are is a prudent check on the inferences one may draw from the behavioral experiments. One would rightly be skeptical of our interpretation of the experiments were it the case that no plausible model could account for the emergence and proliferation of the preferences we have inferred as the proximate explanations of the subjects' behaviors.

Second, the evolutionary models, as we will see, provide insight into the interpretation of the experimental behaviors. The experiments taken alone, do not convincingly distinguish among a number of competing explanations. Evolutionary models can assist in paring down the likely hypotheses for the reconstruction of the behavioral foundations of the social sciences. Indeed evolutionary modeling provides one of the reasons why strong reciprocity is a more convincing interpretation of experimental behavior than competing hypotheses such as inequality aversion and unconditional altruism. Analytical models and agent-based simulations suggest that it is much more likely that plausible learning and inheritance processes would favor strong reciprocity than the other types of social preferences.

Finally, as we will see in the next chapter, the alternative evolutionary models that explain why helping behaviors have become common among humans suggest quite distinct cognitive and affective processes as the proximate causes of helping behavior. These range from the genuine altruism of unconditional maternal love and self-sacrificial loyalty to unrelated com-

rades, to the enlightened self-interest that motivates honoring one's promises in on-going economic exchanges. Thus, the evolutionary origins of cooperative behaviors may hold clues to the psychological states motivating them.