1 Moral Sentiments and Material Interests: Origins, Evidence, and Consequences

Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr

1.1 Introduction

Adam Smith's *The Wealth of Nations* advocates market competition as the key to prosperity. Among its virtues, he pointed out, is that competition works its wonders even if buyers and sellers are entirely self-interested, and indeed sometimes works better if they are. "It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner," wrote Smith, "but from their regard to their own interest" (19). Smith is accordingly often portrayed as a proponent of *Homo economicus*—that selfish, materialistic creature that has traditionally inhabited the economic textbooks. This view overlooks Smith's second—and equally important—contribution, *The Theory of Moral Sentiments*, in which Smith promotes a far more complex picture of the human character.

"How selfish soever man may be supposed," Smith writes in *The Theory of Moral Sentiments*, "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." His book is a thorough scrutiny of human behavior with the goal of establishing that "sympathy" is a central emotion motivating our behavior towards others.

The ideas presented in this book are part of a continuous line of intellectual inheritance from Adam Smith and his friend and mentor David Hume, through Thomas Malthus, Charles Darwin, and Emile Durkheim, and more recently the biologists William Hamilton and Robert Trivers. But Smith's legacy also led in another direction, through David Ricardo, Francis Edgeworth, and Leon Walras, to contemporary neoclassical economics, that recognizes only self-interested behavior.

(V7(M) 24/2/05 16:14) MIT/U J-1147 Gintis AC1:(CKN)21/2/2005 Pal_L (0).27.02.04 (6"×9") pp. 1–40 Ch01_P (p. 3)

The twentieth century was an era in which economists and policy makers in the market economies paid heed only to the second Adam Smith, seeing social policy as the goal of improving social welfare by devising material incentives that induce agents who care only for their own personal welfare to contribute to the public good. In this paradigm, ethics plays no role in motivating human behavior. Albert Hirschman (1985, 10) underscores the weakness of this approach in dealing with crime and corruption:

> Economists often propose to deal with unethical or antisocial behavior by raising the cost of that behavior rather than proclaiming standards and imposing prohibitions and sanctions. . . . [Yet, a] principal purpose of publicly proclaimed laws and regulations is to stigmatize antisocial behavior and thereby to influence citizens' values and behavior codes.

Hirschman argues against a venerable tradition in political philosophy. In 1754, five years before the appearance of Smith's *Theory of Moral Sentiments*, David Hume advised "that, in contriving any system of government . . . every man ought to be supposed to be a knave and to have no other end, in all his actions, than his private interest" (1898 [1754]). However, if individuals are sometimes given to the honorable sentiments about which Smith wrote, prudence recommends an alternative dictum: *Effective policies are those that support socially valued outcomes not only by harnessing selfish motives to socially valued ends, but also by evoking, cultivating, and empowering public-spirited motives*. The research in this book supports this alternative dictum.

We have learned several things in carrying out the research described in this book. First, interdisciplinary research currently yields results that advance traditional intradisciplinary research goals. While the twentieth century was an era of increased disciplinary specialization, the twenty-first may well turn out to be an era of *transdisciplinary synthesis*. Its motto might be: *When different disciplines focus on the same object of knowledge, their models must be mutually reinforcing and consistent where they overlap*. Second, by combining economic theory (game theory in particular) with the experimental techniques of social psychology, economics, and other behavioral sciences, we can empirically test sophisticated models of human behavior in novel ways. The data derived from this unification of disciplinary methods allows us to deduce explicit principles of human behavior that cannot be unambiguously derived using more traditional sources of empirical data.

The power of this experimental approach is obvious: It allows deliberate experimental variation of parameters thought to affect behavior while holding other parameters constant. Using such techniques, experimental economists have been able to estimate the effects of prices and costs on altruistic behaviors, giving precise empirical content to a common intuition that the greater the cost of generosity to the giver and the less the benefit to the recipient, the less generous is the typical experimental subject (Andreoni and Miller 2002).[1] The resulting "supply function of generosity," and other estimates made possible by experiments, are important in underlining the point that other-regarding behaviors do not contradict the fundamental ideas of rationality. They also are valuable in providing interdisciplinary bridges allowing the analytical power of economic and biological models, where other-regarding behavior is a commonly used method, to be enriched by the empirical knowledge of the other social sciences, where it is not.

Because we make such extensive use of laboratory experiments in this book, a few caveats about the experimental method are in order. The most obvious shortcoming is that subjects may behave differently in laboratory and in "real world" settings (Loewenstein 1999). Well-designed experiments in physics, chemistry, or agronomy can exploit the fact that the behavior of entities under study—atoms, agents, soils, and the like—behave similarly whether inside or outside of a laboratory setting. (Murray Gell-Mann once quipped that physics would be a lot harder if particles could think). When subjects *can* think, so-called "experimenter effects" are common. The experimental situation, whether in the laboratory or in the field, is a highly unusual setting that is likely to affect behavioral responses. There is some evidence that experimental behaviors are indeed matched by behaviors in non-experimental settings (Henrich et al. 2001) and are far better predictors of behaviors such as trust than are widely used survey instruments (Glaeser et al. 2000). However, we do not yet have enough data on the behavioral validity of experiments to allay these concerns about experimenter effects with confidence. Thus, while extraordinarily valuable, the experimental approach is not a substitute for more conventional empirical methods, whether statistical, historical, ethnographic, or other. Rather, well-designed experiments may complement these methods. An example, combining behavioral experiments in the field, ethnographic accounts, and cross-cultural statistical hypotheses testing is Henrich et al. 2003.

This volume is part of a general movement toward transdisciplinary research based on the analysis of controlled experimental studies of human behavior, undertaken both in the laboratory and in the field—factories, schools, retirement homes, urban and rural communities, in advanced and in simple societies. Anthropologists have begun to use experimental games as a powerful data instrument in conceptualizing the specificity of various cultures and understanding social variability across cultures (Henrich et al. 2003). Social psychologists are increasingly implementing game-theoretic methods to frame and test hypotheses concerning social interaction, which has improved the quality and interpretability of their experimental data (Hertwig and Ortmann 2001). Political scientists have found similar techniques useful in modeling voter behavior (Frohlich and Oppenheimer 1990; Monroe 1991). Sociologists are finding that analytically modeling the social interactions they describe facilitates their acceptance by scholars in other behavioral sciences (Coleman 1990; Hechter and Kanazawa 1997).

But the disciplines that stand to gain the most from the type of research presented in this volume are economics and human biology. As we have seen, economic theory has traditionally posited that the basic structure of a market economy can be derived from principles that are obvious from casual examination. An example of one of these assumptions is that individuals are *self-regarding*.[2] Two implications of the standard model of self-regarding preferences are in strong conflict with both daily observed preferences and the laboratory and field experiments discussed later in this chapter. The first is the implication that agents care only about the *outcome* of an economic interaction and not about the *process* through which this outcome is attained (e.g., bargaining, coercion, chance, voluntary transfer). The second is the implication that agents care only about what they *personally gain and lose* through an interaction and not what other agents gain or lose (or the nature of these other agents' intentions). Until recently, with these assumptions in place, economic theory proceeded like mathematics rather than natural science; theorem after theorem concerning individual human behavior was proven, while empirical validation of such behavior was rarely deemed relevant and infrequently provided. Indeed, generations of economists learned that the accuracy of its predictions, not the plausibility of its axioms, justifies the neoclassical model of *Homo economicus* (Friedman 1953). Friedman's general position is doubtless defensible, since all tractable models simplify reality. However, we now know that predictions based on the model of the self-

regarding actor often do not hold up under empirical scrutiny, rendering the model inapplicable in many contexts.

A similar situation has existed in human biology. Biologists have been lulled into complacency by the simplicity and apparent explanatory power of two theories: inclusive fitness and reciprocal altruism (Hamilton 1964; Williams 1966; Trivers 1971). Hamilton showed that we do not need amorphous notions of species-level altruism to explain cooperation between related individuals. If a behavior that costs an individual c produces a benefit b for another individual with degree of biological relatedness r (e.g., $r = 0.5$ for parent-child or brother, and $r = 0.25$ for grandparent-grandchild), then the behavior will spread if $r > c/b$. Hamilton's notion of inclusive fitness has been central to the modern, and highly successful, approach to explaining animal behavior (Alcock 1993). Trivers followed Hamilton in showing that even a selfish individual will come to the aid of an unrelated other, provided there is a sufficiently high probability the aid will be repaid in the future. He also was prescient in stressing the fitness-enhancing effects of such seemingly "irrational" emotions and behaviors as guilt, gratitude, moralistic aggression, and reparative altruism. Trivers' reciprocal altruism, which mirrors the economic analysis of exchange between self-interested agents in the absence of costless third-party enforcement (Axelrod and Hamilton 1981), has enjoyed only limited application to nonhuman species (Stephens, McLinn, and Stevens 2002), but became the basis for biological models of human behavior (Dawkins 1976; Wilson 1975).

These theories convinced a generation of researchers that, except for sacrifice on behalf of kin, what appears to be altruism (personal sacrifice on behalf of others) is really just long-run material self-interest. Ironically, human biology has settled in the same place as economic theory, although the disciplines began from very different starting points, and used contrasting logic. Richard Dawkins, for instance, struck a responsive chord among economists when, in *The Selfish Gene* (1989[1976], v.), he confidently asserted "We are survival machines—robot vehicles blindly programmed to preserve the selfish molecules known as genes. . . . This gene selfishness will usually give rise to selfishness in individual behavior." Reflecting the intellectual mood of the times, in his *The Biology of Moral Systems*, R. D. Alexander asserted, "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. . . ." (1987, 3).

The experimental evidence supporting the ubiquity of non–self-regarding motives, however, casts doubt on both the economist's and the biologist's model of the self-regarding human actor. Many of these experiments examine a nexus of behaviors that we term *strong reciprocity*. Strong reciprocity is *a predisposition to cooperate with others, and to punish (at personal cost, if necessary) those who violate the norms of cooperation, even when it is implausible to expect that these costs will be recovered at a later date*.[3] Standard behavioral models of altruism in biology, political science, and economics (Trivers 1971; Taylor 1976; Axelrod and Hamilton 1981; Fudenberg and Maskin 1986) rely on repeated interactions that allow for the establishment of individual reputations and the punishment of norm violators. Strong reciprocity, on the other hand, remains effective even in non-repeated and anonymous situations.[4]

Strong reciprocity contributes not only to the analytical modeling of human behavior but also to the larger task of creating a cogent political philosophy for the twenty-first century. While the writings of the great political philosophers of the past are usually both penetrating and nuanced on the subject of human behavior, they have come to be interpreted simply as having either assumed that human beings are essentially self-regarding (e.g., Thomas Hobbes and John Locke) or, at least under the right social order, entirely altruistic (e.g., Jean Jacques Rousseau, Karl Marx). In fact, people are often neither self-regarding nor altruistic. Strong reciprocators are *conditional cooperators* (who behave altruistically as long as others are doing so as well) and *altruistic punishers* (who apply sanctions to those who behave unfairly according to the prevalent norms of cooperation).

Evolutionary theory suggests that if a mutant gene promotes self-sacrifice on behalf of others—when those helped are unrelated and therefore do not carry the mutant gene and when selection operates only on genes or individuals but not on higher order groups—that the mutant should die out. Moreover, in a population of individuals who sacrifice for others, if a mutant arises that does not so sacrifice, that mutant will spread to fixation at the expense of its altruistic counterparts. Any model that suggests otherwise must involve selection on a level above that of the individual. Working with such models is natural in several social science disciplines but has been generally avoided by a generation of biologists weaned on the classic critiques of group selection by Williams (1966), Dawkins (1976), Maynard Smith (1976), Crow and Kimura (1970), and others, together with the plausible alternatives offered by Hamilton (1964) and Trivers (1971).

But the evidence supporting strong reciprocity calls into question the ubiquity of these alternatives. Moreover, criticisms of group selection are much less compelling when applied to humans than to other animals. The criticisms are considerably weakened when (a) Altruistic punishment is the trait involved and the cost of punishment is relatively low, as is the case for *Homo sapiens*; and/or (b) Either pure cultural selection or gene-culture coevolution are at issue. Gene-culture coevolution (Lumsden and Wilson 1981; Durham 1991; Feldman and Zhivotovsky 1992; Gintis 2003a) occurs when cultural changes render certain genetic adaptations fitness-enhancing. For instance, increased communication in hominid groups increased the fitness value of controlled sound production, which favored the emergence of the modern human larynx and epiglottis. These physiological attributes permitted the flexible control of air flow and sound production, which in turn increased the value of language development. Similarly, culturally evolved norms can affect fitness if norm violators are punished by strong reciprocators. For instance, antisocial men are ostracized in small-scale societies, and women who violate social norms are unlikely to find or keep husbands.

In the case of cultural evolution, the cost of altruistic punishment is considerably less than the cost of unconditional altruism, as depicted in the classical critiques (see chapter 7). In the case of gene-culture coevolution, there may be either no within-group fitness cost to the altruistic trait (although there is a cost to each individual who displays this trait) or cultural uniformity may so dramatically reduce within-group behavioral variance that the classical group selection mechanism—exemplified, for instance, by Price's equation (Price 1970, 1972)—works strongly in favor of selecting the altruistic trait.[5]

Among these models of multilevel selection for altruism is pure genetic group selection (Sober and Wilson 1998), according to which the fitness costs of reciprocators is offset by the tendency for groups with a high fraction of reciprocators to outgrow groups with few reciprocators.[6] Other models involve cultural group selection (Gintis 2000; Henrich and Boyd 2001), according to which groups that transmit a culture of reciprocity outcompete societies that do not. Such a process is as modeled by Boyd, Gintis, Bowles, and Richerson in chapter 7 of this volume, as well as in Boyd et al. 2003. As the literature on the coevolution of genes and culture shows (Feldman, Cavalli-Sforza, and Peck 1985; Bowles, Choi, and Hopfensitz 2003; Gintis 2003a, 2003b), these two alternatives can both be present and mutually reinforcing. These

explanations have in common the idea that altruism increases the fitness of members of groups that practice it by enhancing the degree of cooperation among members, allowing these groups to outcompete other groups that lack this behavioral trait. They differ in that some require *strong* group-level selection (in which the within-group fitness disadvantage of altruists is offset by the augmented average fitness of members of groups with a large fraction of altruists) whereas others require only *weak* group-level selection (in which the within-group fitness disadvantage of altruists is offset by some social mechanism that generates a high rate of production of altruists within the group itself). Weak group selection models such as Gintis (2003a, 2003b) and chapter 4, where supra-individual selection operates only as an equilibrium selection device, avoid the classic problems often associated with strong group selection models (Maynard Smith 1976; Williams 1966; Boorman and Levitt 1980).

This chapter presents an overview of *Moral Sentiments and Material Interests*. While the various chapters of this volume are addressed to readers independent of their particular disciplinary expertise, this chapter makes a special effort to be broadly accessible. We first summarize several types of empirical evidence supporting strong reciprocity as a schema for explaining important cases of altruism in humans. This material is presented in more detail by Ernst Fehr and Urs Fischbacher in chapter 5. In chapter 6, Armin Falk and Urs Fischbacher show explicitly how strong reciprocity can explain behavior in a variety of experimental settings. Although most of the evidence we report is based on behavioral experiments, the same behaviors are regularly observed in everyday life, for example in cooperation in the protection of local environmental public goods (as described by Elinor Ostrom in chapter 9), in wage setting by firms (as described by Truman Bewley in chapter 11), in political attitudes and voter behavior (as described by Fong, Bowles, and Gintis in chapter 10), and in tax compliance (Andreoni, Erard, and Feinstein 1998).

"The Origins of Reciprocity" later in this chapter reviews a variety of models that suggest why, under conditions plausibly characteristic of the early stages of human evolution, a small fraction of strong reciprocators could invade a population of self-regarding types, and a stable equilibrium with a positive fraction of strong reciprocators and a high level of cooperation could result.

While many chapters of this book are based on some variant of the notion of strong reciprocity, Joan Silk's overview of cooperation in

primate species (chapter 2) makes it clear that there are important behavioral forms of cooperation that do not require this level of sophistication. Primates form alliances, share food, care for one another's infants, and give alarm calls—all of which most likely can be explained in terms of long-term self-interest and kin altruism. Such forms of cooperation are no less important in human society, of course, and strong reciprocity can be seen as a generalization of the mechanisms of kin altruism to nonrelatives. In chapter 3, Hillard Kaplan and Michael Gurven argue that human cooperation is an extension of the complex intrafamilial and interfamilial food sharing that is widespread in contemporary hunter-gatherer societies. Such sharing remains important even in modern market societies.

Moreover, in chapter 4, Eric Alden Smith and Rebecca Bliege Bird propose that many of the phenomena attributed to strong reciprocity can be explained in a costly signaling framework. Within this framework, individuals vary in some socially important quality, and higher-quality individuals pay lower marginal signaling costs and thus have a higher optimal level of signaling intensity, given that other members of their social group respond to such signals in mutually beneficial ways. Smith and Bliege Bird summarize an *n*-player game-theoretical signaling model developed by Gintis, Smith, and Bowles (2001) and discuss how it might be applied to phenomena such as provisioning feasts, collective military action, or punishing norm violators. There are several reasons why such signals might sometimes take the form of group-beneficial actions. Providing group benefits might be a more efficient form of broadcasting the signal than collectively neutral or harmful actions. Signal receivers might receive more private benefits from allying with those who signal in group-beneficial ways. Furthermore, once groups in a population vary in the degree to which signaling games produce group-beneficial outcomes, cultural (or even genetic) group selection might favor those signaling equilibria that make higher contributions to mean fitness.

We close this chapter by describing some applications of this material to social policy.

1.2 The Ultimatum Game

In the ultimatum game, under conditions of anonymity, two players are shown a sum of money (say $10). One of the players, called the *proposer*, is instructed to offer any number of dollars, from $1 to $10, to the

second player, who is called the *responder*. The proposer can make only one offer. The responder, again under conditions of anonymity, can either accept or reject this offer. If the responder accepts the offer, the money is shared accordingly. If the responder rejects the offer, both players receive nothing.

Since the game is played only once and the players do not know each other's identity, a self-regarding responder will accept any positive amount of money. Knowing this, a self-regarding proposer will offer the minimum possible amount ($1), which will be accepted. However, when the ultimatum game is actually played, *only a minority of agents behave in a self-regarding manner*. In fact, as many replications of this experiment have documented, under varying conditions and with varying amounts of money, proposers routinely offer respondents very substantial amounts (fifty percent of the total generally being the modal offer), and respondents frequently reject offers below thirty percent (Camerer and Thaler 1995; Güth and Tietz 1990; Roth et al. 1991).

The ultimatum game has been played around the world, but mostly with university students. We find a great deal of individual variability. For instance, in all of the studies cited in the previous paragraph, a significant fraction of subjects (about a quarter, typically) behave in a self-regarding manner. Among student subjects, however, average performance is strikingly uniform from country to country.

Behavior in the ultimatum game thus conforms to the strong reciprocity model: "fair" behavior in the ultimatum game for college students is a fifty-fifty split. Responders reject offers less than forty percent as a form of altruistic punishment of the norm-violating proposer. Proposers offer fifty percent because they are altruistic cooperators, or forty percent because they fear rejection. To support this interpretation, we note that if the offer in an ultimatum game is generated by a computer rather than a human proposer (and if respondents know this), low offers are very rarely rejected (Blount 1995). This suggests that players are motivated by *reciprocity*, reacting to a violation of behavioral norms (Greenberg and Frisch 1972).

Moreover, in a variant of the game in which a responder rejection leads to the responder receiving nothing, but allowing the proposer to keep the share he suggested for himself, respondents never reject offers, and proposers make considerably smaller (but still positive) offers. As a final indication that strong reciprocity motives are operative in this game, after the game is over, when asked why they offer

more than the lowest possible amount, proposers commonly say that they are afraid that respondents will consider low offers unfair and reject them. When respondents reject offers, they usually claim they want to punish unfair behavior.

1.3 Strong Reciprocity in the Labor Market

In Fehr, Gächter, and Kirchsteiger 1997, the experimenters divided a group of 141 subjects (college students who had agreed to participate in order to earn money) into a set of "employers" and a larger set of "employees." The rules of the game are as follows: If an employer hires an employee who provides effort e and receives wage w, his profit is $100e - w$. The wage must be between 1 and 100, and the effort between 0.1 and 1. The payoff to the employee is then $u = w - c(e)$, where $c(e)$ is the "cost of effort" function, which is increasing and convex (the marginal cost of effort rises with effort). All payoffs involve real money that the subjects are paid at the end of the experimental session.

The sequence of actions is as follows. The employer first 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 or her promise, so the employee can choose any effort level, e between .1 and 1 with impunity. Although subjects may play this game several times with different partners, each employer-employee interaction is a one-shot (non-repeated) event. Moreover, the identity of the interacting partners is never revealed.

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, employers will never pay more than the minimum necessary to get the employee to accept a contract, which is 1. The employee will accept this offer, and will set $e = 0.1$. Since $c(0.1) = 0$, the employee's payoff is $u = 1$. The employer's payoff is $(0.1 \times 100) - 1 = 9$.

In fact, however, a majority of agents failed to behave in a self-regarding manner in this experiment.[7] The average net payoff to employees was $u = 35$, and the more generous the employer's wage offer to the employee, the higher the effort the employee provided.

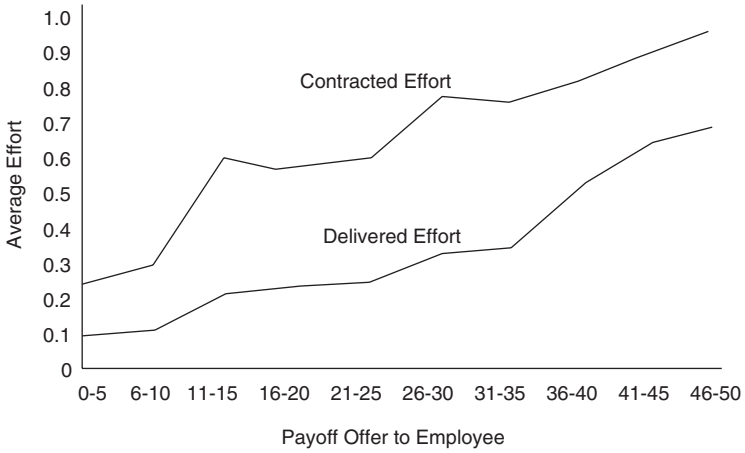

Figure 1.1
Relation of contracted and delivered effort to worker payoff (141 subjects). From Fehr, Gächter, and Kirchsteiger (1997).

In effect, employers presumed the strong reciprocity predispositions of the employees, making quite generous wage offers and receiving higher effort, as a means of increasing both their own and the employee's payoff, as depicted in figure 1.1. Similar results have been observed in Fehr, Kirchsteiger, and Riedl (1993, 1998).

Figure 1.1 also shows that although there is a considerable level of cooperation, there is still a significant gap between the amount of effort agreed upon and the amount actually delivered. This is because, first, only fifty to sixty percent of the subjects are reciprocators, and second, only twenty-six percent of the reciprocators delivered the level of effort they promised! We conclude that strong reciprocators are inclined to compromise their morality to some extent.

This evidence is compatible with the notion that the employers are purely self-regarding, since their beneficent behavior vis-à-vis their employees was effective in increasing employer profits. To see if employers are also strong reciprocators, the authors extended the game following the first round of experiments by allowing the employers to respond reciprocally 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 self-regarding, they would of course do neither, since they would not interact with the same worker a second time. However, sixty-eight percent of the time employers punished

employees that did not fulfill their contracts, and seventy percent of the time employers rewarded employees who overfulfilled their contracts. Indeed, employers rewarded forty-one percent 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. Underfulfilling contracts dropped from eighty-three to twenty-six percent of the exchanges, and overfulfilled contracts rose from three to thirty-eight percent of the total. Finally, allowing employers to reward and punish led to a forty-percent increase in the net payoffs to all subjects, even when the payoff reductions resulting from employer punishment of employees are taken into account.

We conclude from this study that the subjects who assume the role of employee conform to internalized standards of reciprocity, 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 are rewarded for acting accordingly. Finally, employers draw upon the internalized norm of rewarding good and punishing bad behavior when they are permitted to punish, and employees expect this behavior and adjust their own effort levels accordingly.

1.4 The Public Goods Game

The *public goods game* has been analyzed in a series of papers by the social psychologist Toshio Yamagishi (1986, 1988a, 1998b), by the political scientist Elinor Ostrom and her coworkers (Ostrom, Walker, and Gardner 1992), and by economists Ernst Fehr and his coworkers (Gächter and Fehr 1999; Fehr and Gächter 2000a, 2002). These researchers uniformly found that *groups exhibit a much higher rate of cooperation than can be expected assuming the standard model of the self-regarding actor*, and this is especially the case when subjects are given the option of incurring a cost to themselves in order to punish free-riders.

A typical public goods game has several rounds, say ten. The subjects are told the total number of rounds and all other aspects of the game and are paid their winnings in real money at the end of the session. In each round, each subject is grouped with several other subjects—say three others—under conditions of strict anonymity. Each subject is then given a certain number of "points," say twenty,

redeemable at the end of the experimental session for real money. Each subject then places some fraction of his points in a "common account" and the remainder in the subject's own "private account."

The experimenter then tells the subjects how many points were contributed to the common account and adds to the private account of each subject some fraction of the total amount in the common account, say forty percent. So if a subject contributes his or her whole twenty points to the common account, each of the four group members will receive eight points at the end of the round. In effect, by putting her or his whole endowment into the common account, a player loses twelve points but the other three group members gain a total of twenty-four ($= 8 \times 3$) points. The players keep whatever is in their private accounts at the end of each round.

A self-regarding player will contribute nothing to the common account. However, only a fraction of subjects in fact conform to the self-interest model. Subjects begin by contributing on average about half of their endowments to the public account. The level of contributions decays over the course of the ten rounds, until in the final rounds most players are behaving in a self-regarding manner (Dawes and Thaler 1988; Ledyard 1995). In a metastudy of twelve public goods experiments, Fehr and Schmidt (1999) found that in the early rounds, average and median contribution levels ranged from forty to sixty percent of the endowment, but in the final period seventy-three percent of all individuals ($N = 1042$) contributed nothing, and many of the other players contributed close to zero. These results are not compatible with the selfish-actor model (which predicts zero contribution in all rounds), although they might be predicted by a reciprocal altruism model, since the chance to reciprocate declines as the end of the experiment approaches.

However this is not in fact the explanation of the moderate but deteriorating levels of cooperation in the public goods game. The subjects' own explanation of the decay of cooperation after the experiment is that cooperative subjects became angry with others who contributed less than themselves and retaliated against free-riding low contributors in the only way available to them—by lowering their own contributions (Andreoni 1995).

Experimental evidence supports this interpretation. When subjects are allowed to punish noncontributors, they do so at a cost to themselves (Orbell, Dawes, and Van de Kragt 1986; Sato 1987; Yamagishi 1988a, 1988b, 1992). For instance, in Ostrom, Walker, and Gardner

(1992), subjects interacted for twenty-five periods in a public goods game. By paying a "fee," subjects could impose costs on other subjects by "fining" them. Since fining costs the individual who uses it, and the benefits of increased compliance accrue to the group as a whole, assuming agents are self-regarding, no player ever pays the fee, no player is ever punished for defecting, and all players defect by contributing nothing to the common pool. However, the authors found a significant level of punishing behavior in this version of the public goods game.

These experiments allowed individuals to engage in strategic behavior, since costly punishment of defectors could increase cooperation in future periods, yielding a positive net return for the punisher. Fehr and Gächter (2000a) set up an experimental situation in which the possibility of strategic punishment was removed. They employed three different methods of assigning study subjects to groups of four individuals each. The groups played six- and ten-round public goods games with costly punishment allowed at the end of each round. There were sufficient subjects to run between ten and eighteen groups simultaneously. Under the *partner treatment*, the four subjects remained in the same group for all ten rounds. Under the *stranger treatment*, the subjects were randomly reassigned after each round. Finally, under the *perfect stranger treatment*, the subjects were randomly reassigned and assured that they would never meet the same subject more than once.

Fehr and Gächter (2000a) performed their experiment over ten rounds with punishment and then over ten rounds without punishment.[8] Their results are illustrated in figure 1.2. We see that when costly punishment is permitted, cooperation does not deteriorate, and in the partner game, despite strict anonymity, cooperation increases to almost full cooperation, even in the final round. When punishment is not permitted, however, the same subjects experience the deterioration of cooperation found in previous public goods games. The contrast in cooperation rates between the partner and the two stranger treatments is worth noting, because the strength of punishment is roughly the same across all treatments. This suggests that the credibility of the punishment threat is greater in the partner treatment because the punished subjects are certain that, once they have been punished in previous rounds, the punishing subjects remain in their group. The impact of strong reciprocity on cooperation is thus more strongly manifested when the group is the more coherent and permanent.

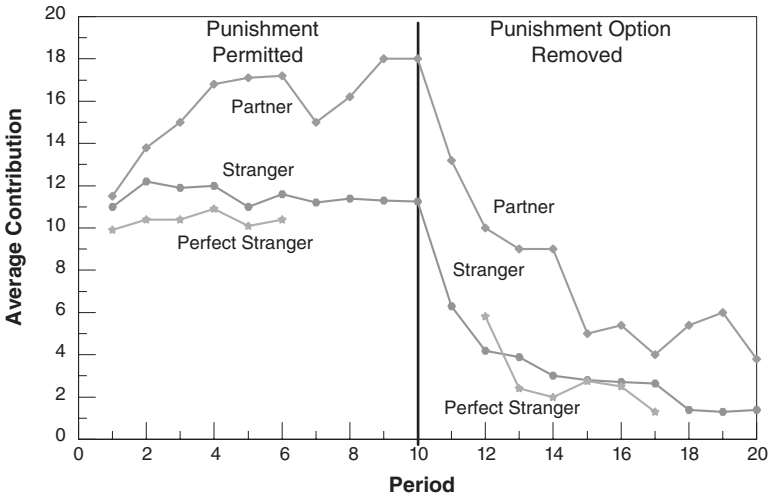

Figure 1.2
Average contributions over time in the partner, stranger, and perfect stranger treatments when the punishment condition is played first. Adapted from Fehr and Gächter 2000a.

1.5 Intentions or Outcomes?

One key fact missing from the discussion of public goods games is a specification of the relationship between contributing and punishing. The strong reciprocity interpretation suggests that high contributors will be high punishers and punishees will be below-average contributors. This prediction is borne out in Fehr and Gächter (2002), where seventy-five percent of the punishment acts carried out by the 240 subjects were executed by above-average contributors, and the most important variable in predicting how much one player punished another was the difference between the punisher's contribution and the punishee's contribution.

Another key question in interpreting public goods games is: Do reciprocators respond to fair or unfair *intentions* or do they respond to fair or unfair *outcomes*? The model of strong reciprocity unambiguously favors intentions over outcomes. To answer this question, Falk, Fehr, and Fischbacher (2002) ran two versions of the "moonlighting game"—an intention treatment (I-treatment) where a player's intentions could be deduced from his action, and a no-intention treatment (NI-treatment), where a player's intentions could not be deduced. They provide clear and unambiguous evidence for the behavioral rele-

vance of intentions in the domain of both negatively and positively reciprocal behavior.

The moonlighting game consists of two stages. At the beginning of the game, both players are endowed with twelve points. At the first stage player A chooses an action a in $\{-6, -5, \ldots, 5, 6\}$. If A chooses $a > 0$, he gives player B a tokens, while if he chooses $a < 0$, he takes away $|a|$ tokens from B. In case $a \geq 0$, the experimenter triples a so that B receives $3a$. After B observes a, he can choose an action b in $\{-6, -5, \ldots 17, 18\}$. If $b \geq 0$, B gives the amount b to A. If $b < 0$, B loses $|b|$, and A loses $|3b|$. Since A can give and take while B can reward or sanction, this game allows for both positively and negatively reciprocal behavior. Each subject plays the game only once.

If the Bs are self-regarding, they will all choose $b = 0$, neither rewarding nor punishing their A partners, since the game is played only once. Knowing this, if the As are self-regarding, they will all choose $a = -6$, which maximizes their payoff. In the I-treatment, A players are allowed to choose a, whereas in the NI-treatment, A's choice is determined by a roll of a pair of dice. If the players are not self-regarding and care only about the fairness of the outcomes and not intentions, there will be no difference in the behavior of the B players across the I- and the NI-treatments. Moreover, if the A players believe their B partners care only about outcomes, their behavior will not differ across the two treatments. If the B players care only about the intentions of their A partners, they will never reward or punish in the NI-treatment, but they will reward partners who choose high $a > 0$ and punish partners who choose $a < 0$.

The experimenters' main result was that the behavior of player B in the I-treatment is substantially different from the behavior in the NI-treatment, indicating that the attribution of fairness intentions is behaviorally important. Indeed, As who gave to Bs were generally rewarded by Bs in the I-treatment much more that in the NI-treatment (significant at the 1 level), and As who took from Bs were generally punished by Bs in the I-treatment much more than in the NI-treatment (significant at the 1 level).

Turning to individual patterns of behavior, in the I-treatment, no agent behaved purely selfishly (i.e., no agent set $b = 0$ independent of a), whereas in the NI-treatment thirty behaved purely selfishly. Conversely, in the I-treatment seventy-six percent of subjects rewarded or sanctioned their partner, whereas in the NI-treatment, only thirty-nine percent of subjects rewarded or sanctioned. We conclude that most

agents are motivated by the intentionality of their partners, but a significant fraction care about the outcome, either exclusively or in addition to the intention of the partner.

1.6 Crowding Out

There are many circumstances in which people voluntarily engage in an activity, yet when monetary incentives are added in an attempt to increase the level of the activity, the level actually decreases. The reason for this phenomenon, which is called *crowding out*, is that the number of contributors responding to the monetary incentives is more than offset by the number of discouraged voluntary contributors. This phenomenon was first stressed by Titmuss (1970), noting that voluntary blood donation in Britain declined sharply when a policy of paying donors was instituted alongside the voluntary sector. More recently, Frey (1997a, 1997b, 1997c) has applied this idea to a variety of situations. In chapter 9 of this volume, Elinor Ostrom provides an extremely important example of crowding out. Ostrom reviews the extensive evidence that when the state regulates common property resources (such as scare water and depletable fish stocks) by using fines and subsidies to encourage conservation, the overuse of these resources may actually increase. This occurs because the voluntary, community-regulated, system of restraints breaks down in the face of relatively ineffective formal government sanctions.

In many cases, such crowding out can be explained in a parsimonious manner by strong reciprocity. Voluntary behavior is the result of what we have called the *predisposition to contribute to a cooperative endeavor*, contingent upon the cooperation of others. The monetary incentive to contribute destroys the cooperative nature of the task, and the threat of fining defectors may be perceived as being an unkind or hostile action (especially if the fine is imposed by agents who have an antagonistic relationship with group members). The crowding out of voluntary cooperation and altruistic punishment occur because the preconditions for the operation of strong reciprocity are removed when explicit material incentives are applied to the task.

This interpretation is supported by the laboratory experiment of Fehr and Gächter (2000b), who show that in an employer–employee setting (see Strong Reciprocity in the Labor Market) if an employer explicitly threatens to fine a worker for malfeasance, the worker's willingness to cooperate voluntarily is significantly reduced. Similarly,

Fehr and List (2002) report that chief executive officers respond in a less trustworthy manner if they face a fine compared to situations where they do not face a fine.

As a concrete example, consider Fehr and Rockenbach's (2002) experiment involving 238 subjects. Mutually anonymous subjects are paired, one subject having the role of *investor*, the other *responder*. They then play a *trust game* in which both subjects receive ten money units (MUs). The investor can transfer any portion of his endowment to the responder and must specify a *desired return* from the responder, which could be any amount less than or equal to what the responder receives as a result of tripling the investor's transfer. The responder, knowing both the amount sent and the amount the investor wants back, chooses an amount to send back to the investor (not necessarily the amount investor requested). The investor receives this amount (which is not tripled), and the game is over.

There were two experimental conditions—a *trust* condition with no additional rules and an *incentive* condition that adds one more rule: the investor has the option of precommitting to impose a fine of four MUs on the responder should the latter return less than the investor's desired return. At the time the investor chooses the transfer and the desired return, he also must specify whether to impose the fine condition. The responder then knows the transfer, the desired return, and whether the fine condition was imposed by the investor.

Since all the interactions in this game are anonymous and there is only one round, self-regarding respondents will return nothing in the trust condition and at most four MUs in the incentive condition. Thus, self-regarding investors who expect their partners to be self-regarding will send nothing to responders in the trust condition and will not ask for more than four MUs back in the incentive condition. Assuming a respondent will only avoid the fine if he can gain from doing so, the investor will transfer two MUs and ask for three MUs back, the responder will get six MUs and return three MUs to the investor. It follows that if all agents are self-regarding and all know that this is the case, investors will always choose to impose the fine condition and end up with eleven MUs, while the responders end up with thirteen MUs.

In contrast to this hypothesis, responders actually paid back substantial amounts of money under all conditions. In addition, responders' returns to investors were highest when the investor *refrained* from imposing the fine in the incentive condition and were lowest

when the investor imposed the fine condition in the incentive condition. Returns were intermediate under the trust condition where fines could not be imposed.

The experimenters ascertained that the greater return when the fine was not imposed could not be explained either by investors in that situation transferring more to the responders or by investors requesting more modest returns from the respondents. But if we assume that imposing the fine condition is interpreted as a hostile act by the respondent, and hence not imposing this condition is interpreted as an act of kindness and trust, then strong reciprocity supplies a plausible reason why responders increase their compliance with investors' requests when the investors refrain from fining them.

1.7 The Origins of Strong Reciprocity

Some behavioral scientists, including many sociologists and anthropologists, are quite comfortable with the notion that altruistic motivations are an important part of the human repertoire and explain their prevalence by cultural transmission. Support for a strong cultural element in the expression of both altruistic cooperation and punishment can be drawn from the wide variation in strength of both cooperation and punishment exhibited in our small-scale societies study (Henrich et al. [2001] and this chapter's discussion of the ultimatum game), and our ability to explain a significant fraction of the variation in behavior in terms of social variables (cooperation in production and degree of market integration). Even though altruists must bear a fitness cost for their behavior not shared by self-regarding types, in most cases this cost is not high—shunning, gossip, and ostracism, for instance (Bowles and Gintis 2004). Indeed, as long as the cultural system transmits altruistic values strongly enough to offset the fitness costs of altruism, society can support motivations that are not fitness-maximizing indefinitely (Boyd and Richerson 1985; Gintis 2003b). Moreover, societies with cultural systems that promote cooperation will outcompete those that do not, and individuals tend to copy the behaviors characteristic of successful groups. Together, these forces can explain the diffusion of group-beneficial cultural practices (Soltis, Boyd, and Richerson 1995; Boyd and Richerson 2002).

While culture is part of the explanation, it is possible that strong reciprocity, like kin altruism and reciprocal altruism, has a significant genetic component. Altruistic punishment, for instance, is not cultur-

ally transmitted in many societies where people regularly engage in it (Brown 1991). In the Judeo-Christian tradition, for example, charity and forgiveness ("turn the other cheek") are valued, while seeking revenge is denigrated. Indeed, willingness to punish transgressors is not seen as an admirable personal trait and, except in special circumstances, people are not subject to social opprobrium for failing to punish those who hurt them.

If this is the case, the altruistic behaviors documented and modeled in this book indicate that gene-culture coevolution has been operative for human beings. This is indeed what we believe to be the case, and in this section we describe some plausible coevolutionary models that could sustain strong reciprocity. It is thus likely that strong reciprocity is the product of gene-culture coevolution. It follows that group level-characteristics that enhance group selection pressures—such as relatively small group size, limited migration, or frequent intergroup conflicts—coevolved with cooperative behaviors. This being the case, we concluded that cooperation is based in part on the distinctive capacities of humans to construct institutional environments that limit within-group competition and reduce phenotypic variation within groups, thus heightening the relative importance of between-group competition and allowing individually-costly but ingroup-beneficial behaviors to coevolve within these supporting environments through a process of interdemic group selection.

The idea that the suppression of within-group competition may be a strong influence on evolutionary dynamics has been widely recognized in eusocial insects and other species. Boehm (1982) and Eibl-Eibesfeldt (1982) first applied this reasoning to human evolution, exploring the role of culturally transmitted practices that reduce phenotypic variation within groups. Examples of such practices are leveling institutions, such as monogamy and food sharing among nonkin (namely, those practices which reduce within-group differences in reproductive fitness or material well-being). By reducing within-group differences in individual success, such structures may have attenuated within-group genetic or cultural selection operating against individually-costly but group-beneficial practices, thus giving the groups adopting them advantages in intergroup contests. Group-level institutions are thus constructed environments capable of imparting distinctive direction and pace to the process of biological evolution and cultural change. Hence, the evolutionary success of social institutions that reduce phenotypic variation within groups may be explained by the

fact that they retard selection pressures working against ingroup–beneficial individual traits and that high frequencies of bearers of these traits reduces the likelihood of group extinctions (Bowles, Choi, and Hopfensitz 2003).

In chapter 8, Rajiv Sethi and E. Somanathan provide an overview of evolutionary models of reciprocity conforming to the logic described in the previous paragraph and also present their own model of common property resource use. In their model, there are two types of individuals: *reciprocators* who choose extraction levels that are consistent with efficient and fair resource use, monitor other users, and punish those who over-extract relative to the norm; and *opportunists* who choose their extraction levels optimally in response to the presence or absence of reciprocators and do not punish. Since monitoring is costly, and opportunists comply with the norm only when it is in their interest to do so, reciprocators obtain lower payoffs than opportunists within all groups, regardless of composition. However, since the presence of reciprocators alters the behavior of opportunists in a manner that benefits all group members, a population of opportunists can be unstable under random (non-assortative) matching. More strikingly, even when a population of opportunists is stable, Sethi and Somanathan show that stable states in which a mix of reciprocators and opportunists is present can exist.

In chapter 7, Robert Boyd, Herbert Gintis, Samuel Bowles, and Peter J. Richerson explore a deep asymmetry between altruistic cooperation and altruistic punishment. They show that altruistic punishment allows cooperation in quite large groups because the payoff disadvantage of altruistic cooperators relative to defectors is independent of the frequency of defectors in the population, while the cost disadvantage of those engaged in altruistic punishment declines as defectors become rare. Thus, when altruistic punishers are common, selection pressures operating against them are weak. The fact that punishers experience only a small disadvantage when defectors are rare means that weak within-group evolutionary forces, such as conformist transmission, can stabilize punishment and allow cooperation to persist. Computer simulations show that selection among groups leads to the evolution of altruistic punishment when it could not maintain altruistic cooperation without such punishment.

The interested reader will find a number of related cultural and gene-culture coevolution models exhibiting the evolutionary stability of altruism in general, and strong reciprocity in particular, in recent

papers (Gintis 2000; Bowles 2001; Henrich and Boyd 2001; and Gintis 2003a).

1.8 Strong Reciprocity: Altruistic Adaptation or Self-Interested Error?

There is an alternative to our treatment of altruistic cooperation and punishment that is widely offered in reaction to the evidence upon which our model of strong reciprocity is based. The following is our understanding of this argument, presented in its most defensible light.

Until about 10,000 years ago—before the advent of sedentary agriculture, markets, and urban living—humans were generally surrounded by kin and long-term community consociates. Humans were thus rarely called upon to deal with strangers or interact in one-shot situations. During the formative period in our evolutionary history, therefore, humans developed a cognitive and emotional system that reinforces cooperation among extended kin and others with whom one lives in close and frequent contact, but developed little facility for behaving differently when facing strangers in non-repeatable and/or anonymous settings. Experimental games therefore confront subjects with settings to which they have not evolved optimal responses. It follows that strong reciprocity is simply irrational and mistaken behavior. This accounts for the fact that the same behavior patterns and their emotional correlates govern subject behavior in both anonymous, one-shot encounters and when subjects' encounters with kin and long-term neighbors. In sum, strong reciprocity is an historically evolved form of enlightened self- and kin-interest that falsely appears altruistic when deployed in social situations for which it was not an adaptation.

From an operational standpoint, it matters little which of these views is correct, since human behavior is the same in either case. However, if altruism is actually misapplied self-interest, we might expect altruistic behavior to be driven out of existence by consistently self-regarding individuals in the long run. If these arguments are correct, it would likely lead to the collapse of the sophisticated forms of cooperation that have arisen in civilized societies. Moreover, the alternative suggests that agents can use their intellect to "learn" to behave selfishly when confronted with the results of their suboptimal behavior. The evidence, however, suggests that cooperation based on strong reciprocity can

unravel when there is no means of punishing free-riders but that it does not unravel simply through repetition.

What is wrong with the alternative theory? First, it is probably not true that prehistoric humans lived in groups comprised solely of close kin and long-term neighbors. Periodic social crises in human prehistory, occurring at roughly thirty-year intervals on average, are probable, since population contractions were common (Boone and Kessler 1999) and population crashes occurred in foraging groups at a mean rate of perhaps once every thirty years (Keckler 1997). These and related archaeological facts suggest that foraging groups had relatively short lifespans.

If the conditions under which humans emerged are similar to the conditions of modern primates and/or contemporary hunter-gatherer societies, we can reinforce our argument by noting that there is a constant flow of individuals into and out of groups in such societies. Exogamy alone, according to which young males or females relocate to other groups to seek a mate, gives rise to considerable intergroup mixing and frequent encounters with strangers and other agents with whom one will not likely interact in the future. Contemporary foraging groups, who are probably not that different in migratory patterns from their prehistoric ancestors, are remarkably outbred compared to even the simplest farming societies, from which we can infer that dealing with strangers in short-term relationships was a common feature of our evolutionary history. Henry Harpending (email communication) has found in his studies of the Bushmen in the Kalahari that there were essentially random patterns of mating over hundreds of kilometers. See Fix (1999) for an overview and analysis of the relevant data on this issue.

Second, if prehistoric humans rarely interacted with strangers, then our emotional systems should not be finely tuned to degrees of familiarity—we should treat all individuals as neighbors. But we in fact are quite attuned to varying degrees of relatedness and propinquity. Most individuals care most about their children, next about their close relatives, next about their close neighbors, next about their conationals, and so on, with decreasing levels of altruistic sentiment as the bonds of association grow weaker. Even in experimental games, repetition and absence of anonymity dramatically increase the level of cooperation and punishment. There is thus considerable evidence that altruistic cooperation and punishment in one-shot and anonymous settings is the product of evolution and not simply errant behavior.

1.9 Strong Reciprocity and Cultural Evolution

Strong reciprocity is a *behavioral schema* that is compatible with a wide variety of cultural norms. Strong reciprocators are predisposed to cooperate in social dilemmas, but the particular social situations that will be recognized as appropriate for cooperation are culturally variable. Strong reciprocators punish group members who behave selfishly, but the norms of fairness and the nature of punishment are culturally variable.

In this section, we first present evidence that a wide variety of cultural forms are compatible with strong reciprocity. We then argue that the strong reciprocity schema is capable of stabilizing a set of cultural norms, whether or not these norms promote the fitness of group members. Finally, we suggest that the tendency for strong reciprocity to be attached to prosocial norms can be accounted for by intergroup competition, through which societies prevail over their competitors to the extent that their cultural systems are fitness enhancing.

1.9.1 Cultural Diversity

What are the limits of cultural variability, and how does strong reciprocity operate in distinct cultural settings? To expand the diversity of cultural and economic circumstances of experimental subjects, we undertook a large cross-cultural study of behavior in various games including the ultimatum game (Henrich et al. 2001; Henrich et al. 2003). Twelve experienced field researchers, working in twelve countries on four continents, recruited subjects from fifteen small-scale societies exhibiting a wide variety of economic and cultural conditions. These societies consisted of three foraging groups (the Hadza of East Africa, the Au and Gnau of Papua New Guinea, and the Lamalera of Indonesia), six slash-and-burn horticulturists and agropasturalists (the Aché, Machiguenga, Quichua, Tsimané, and Achuar of South America, and the Orma of East Africa), four nomadic herding groups (the Turguud, Mongols, and Kazakhs of Central Asia, and the Sangu of East Africa) and two sedentary, small-scale agricultural societies (the Mapuche of South America and Zimbabwean farmers in Africa).

We can summarize our results as follows. First, the canonical model of self-regarding behavior is not supported in *any* of the societies studied. In the ultimatum game, for example, in all societies either responders, proposers, or both behaved in a reciprocal manner. Second, there is considerably more behavioral variability across groups than

had been found in previous cross-cultural research. While mean ultimatum game offers in experiments with student subjects are typically between forty-three and forty-eight percent, the mean offers from proposers in our sample ranged from twenty-six to fifty-eight percent. While modal ultimatum game offers are consistently fifty percent among university students, sample modes with the data range in this study ranged from fifteen to fifty percent. Rejections were extremely rare, in some groups (even in the presence of very low offers), while in others, rejection rates were substantial, including frequent rejections of *hyper-fair* offers (i.e., offers above fifty percent). By contrast, the Machiguenga have mean offer of twenty-six percent but no rejections. The Aché and Tsimané distributions resemble inverted American distributions. The Orma and Huinca (non–Mapuche Chileans living among the Mapuche) have modes near the center of the distribution, but show secondary peaks at full cooperation.

Third, *differences between societies in "market integration" and "cooperation in production" explain a substantial portion (about fifty percent) of the behavioral variation between groups*. The higher the degree of market integration and the higher the payoffs to cooperation, the greater the level of cooperation and sharing in experimental games. The societies were rank-ordered in five categories—market integration (how often do people buy and sell, or work for a wage?), cooperation in production (is production collective or individual?), plus anonymity (how prevalent are anonymous roles and transactions?), privacy (how easily can people keep their activities secret?), and complexity (how much centralized decision-making occurs above the level of the household?). Using statistical regression analysis, only the first two characteristics were significant, and they together accounted for about fifty percent of the variation among societies in mean ultimatum game offers. Fourth, individual-level economic and demographic variables do not explain behavior either within or across groups. Finally, the nature and degree of cooperation and punishment in the experiments is generally consistent with economic patterns of everyday life in these societies.

The final point of this experiment is in some respects the most important for future research. In a number of cases, the parallels between experimental game play and the structure of daily life were quite striking. Nor was this relationship lost on the subjects themselves. The Orma immediately recognized that the public goods game was similar to the *harambee*, a locally-initiated contribution that households make when a community decides to construct a road or school. They dubbed

the experiment "the harambee game" and gave generously (mean fifty-eight percent with twenty-five percent full contributors).

Among the Au and Gnau, many proposers offered more than half the total amount and many of these hyper-fair offers were rejected! This reflects the Melanesian culture of status-seeking through gift giving. Making a large gift is a bid for social dominance in everyday life in these societies, and rejecting the gift is a rejection of being subordinate.

Among the whale-hunting Lamalera, sixty-three percent of the proposers in the ultimatum game divided the total amount equally, and most of those who did not offered more than fifty percent (the mean offer was fifty-seven percent). In real life, a large catch—always the product of cooperation among many individual whalers—is meticulously divided into predesignated proportions and carefully distributed among the members of the community.

Among the Aché, seventy-nine percent of proposers offered either forty or fifty percent, and sixteen percent offered more than fifty percent, with no rejected offers. In daily life, the Aché regularly share meat, which is distributed equally among all households irrespective of which hunter made the catch.

In contrast to the Aché, the Hadza made low offers and had high rejection rates in the ultimatum game. This reflects the tendency of these small-scale foragers to share meat but with a high level of conflict and frequent attempts of hunters to hide their catch from the group.

Both the Machiguenga and Tsimané made low ultimatum game offers, and there were virtually no rejections. These groups exhibit little cooperation, exchange, or sharing beyond the family unit. Ethnographically, both groups show little fear of social sanctions and care little about "public opinion."

The Mapuche's social relations are characterized by mutual suspicion, envy, and fear of being envied. This pattern is consistent with researchers' interviews with the Mapuche following the ultimatum game. Mapuche proposers rarely claimed that their offers were influenced by fairness but rather by a fear of rejection. Even proposers who made hyper-fair offers claimed that they feared the remote possibility of spiteful responders, who would be willing to reject even fifty-fifty offers.

1.9.2 Cultural Evolution

Suppose a group, in the name of promoting group harmony, has adopted the norm of peaceful adjudication of disputes. If the members

are self-interested, no third party will intervene in a dispute between two members to thwart a violent interaction and punish its perpetrators. By contrast, a group with a sufficient fraction of reciprocators will intervene, allowing the norm to persist over time, even in the face of the indifference of the self-interested and the opposition of an appreciable fraction of troublemakers. Thus, strong reciprocity can stabilize prosocial norms that otherwise could not be sustained in the group.

Conversely, suppose in the name of preventing invidious distinctions, a group has adopted a work norm that discourages members from supplying effort above a certain approved level. Such a norm is, of course, fitness-reducing for the group's members. Indeed, if members are self-interested, some will violate the norm, and no others will intervene to protect it. The fitness-reducing norm will thus disappear. However, a small fraction of strong reciprocators who accept the norm and who punish its violators can stabilize the norm even when many would prefer to violate it.

Our point here is simple. For most of human history (until a few thousand years ago), there were no schools, churches, books, laws, or states. There was, therefore, no centralized institutional mechanism for enforcing norms that affect the members of a group as a whole. Strong reciprocity evolved because groups with strong reciprocators were capable of stabilizing prosocial norms that could not be supported using principles of long-term self-interest alone, because it is generally fitness-enhancing for an individual to punish only transgressions against the individual himself and then only if the time horizon is sufficiently lengthy to render a reputation for protecting one's interests. On the other hand, the same mechanisms that have the ability to enforce prosocial norms can almost as easily enforce fitness-neutral and antisocial norms (Edgerton 1992; Boyd and Richerson 1992; Richerson and Boyd 2003).

In this framework, prosocial norms evolve not because they have superior fitness within groups, but because groups with prosocial norms outcompete groups that are deficient in this respect. It is not surprising, for instance, that the "great religions" (Judaism, Christianity, Buddhism, Islam, Hinduism, and so forth) stress prosocial norms—such as helping one's neighbors, giving each his due, turning the other cheek, and the like.

There is considerable evidence for the operation of natural selection in cultural evolution (Richerson and Boyd 2003). For instance, religious practice differences entail fertility and survival differentials (Roof and

McKinney 1987), and the organization of human populations into units which engage in sustained, lethal combat with other groups leads to the survival of groups with prosocial organizational and participatory forms. Soltis, Boyd, and Richerson (1995) reviewed the ethnography of warfare in simple societies in highland New Guinea. The pattern of group extinction and new group formation in these cases conforms well to a cultural evolution model. The strength of cultural group selection in highland New Guinea was strong enough to cause the spread of a favorable new social institution among a metapopulation in about 1,000 years. Cases of group selection by demic expansion are quite well described, for example the spread of the southern Sudanese Nuer at the expense of the Dinka (Kelly 1985), the expansion of the Marind-anim at the expense of their neighbors by means of large, well-organized head-hunting raids at the expense of their neighbors, including the capture and incorporation of women and children (Knauft 1993), and the Hispanic conquest of Latin America (Foster 1960).

1.10 Applications to Social Policy

Economic policy has generally been based on a model of the self-regarding individual. It would be surprising if our model of strong reciprocity did not suggest significant revisions in standard economic policy reasoning, and indeed it does. This section includes several applications of the strong reciprocity model to social policy. In fact, only a relatively weak version of strong reciprocity enters into policy analysis. All that is required is that agents be conditional cooperators and altruistic punishers in public and repeated situations where reputations can be established—an assumption amply justified by the behavioral evidence. Specifically, it is unimportant for these analyses whether strong reciprocity is the product of purely cultural or gene-culture coevolutionary dynamics—whether this behavior is truly altruistic or includes some difficult-to-observe personal payoff (such as costly signaling, as suggested by Smith and Bliege Bird in chapter 4), or whether it is fundamentally adaptive or maladaptive.

Elinor Ostrom argues in chapter 9 that common pool resource management has often failed when based on the standard model of incentives, whereas a more balanced program of local community management and government regulation—often the former alone—can contribute to effective conservation and egalitarian distribution of

common pool resources. This alternative policy framework flows naturally from the strong reciprocity model and depends on the presence of a fraction of strong reciprocators in the population for its effectiveness.

As Christina Fong, Samuel Bowles, and Herbert Gintis show in chapter 10, approaches to egalitarian income redistribution are also strengthened by the use of the strong reciprocity model. During the last few decades of the twentieth century in the United States, there emerged an unprecedented malaise concerning the system of egalitarian redistribution in public opinion. Many interpret this shift, which has led to important changes in the social welfare system, as a resurgence of self-interest on the part of the country's nonpoor and of racist attitudes on the part of the majority white citizenry. Fong, Bowles, and Gintis present a body of evidence that disputes this view and argue in favor of model of voter behavior based on strong reciprocity.

In chapter 11, Truman Bewley uses strong reciprocity to model unemployment in the macroeconomy of the United States. Bewley tackles one of the oldest, and most controversial, puzzles in economics: why nominal wages rarely fall (and real wages do not fall enough) when unemployment is high. He does so in a novel way, through interviews with over 300 businessmen, union leaders, job recruiters, and unemployment counselors in the northeastern United States during the early 1990s recession. Bewley concludes that employers resist pay cuts largely because the savings from lower wages are usually outweighed by the cost of reducing worker morale: pay cuts are seen by workers as an unfriendly and unfair act, and employees retaliate by working less hard and less in line with managements' goals. Bewley thus shows that even the most standard of economic problems, that of wage determination, cannot be understood outside the framework of an empirical and behavioral approach to individual behavior.

Nowhere has the standard model of the self-regarding actor had more influence than in legal theory and the politics of legislation. Beginning with the work of economist Ronald Coase (1960) and developed by the legal scholar Richard Posner (1973), "Law and Economics" has become a potent analytical framework for studying the effect of legislation on social welfare. While we do not doubt the value of this work, its abstraction from reciprocity and other non–self-regarding motives limits its general relevance. In chapter 12, Dan M. Kahan addresses the relevance of reciprocity to law and public policy. He suggests that individuals will often contribute voluntarily to collective goods so long as they believe that most others are willing to do the

same. Promoting trust, in the form of reason to believe that fellow citizens are contributing their fair share, is thus a potential alternative to costly incentive schemes for solving societal collective action problems. Indeed, conspicuous penalties and subsidies, reciprocity theory implies, might sometimes aggravate rather than ameliorate collective action problems by giving citizens reason to doubt that other citizens are contributing voluntarily to societal collective goods. He illustrates these conclusions by analyzing several regulatory problems—including tax evasion, the location of toxic waste facilities, and the production of information and technology.

In the final chapter of this volume, Samuel Bowles and Herbert Gintis offer a larger and more synthetic vision of what a deeper appreciation of moral sentiments might imply for social structure and policy. They argue that the moral sentiments documented and analyzed in this book lead us to a new view of social communities and an understanding of why the two preeminently anonymous modern institutions—the market and the state—only incompletely addresses modern social problems.

If Bowles and Gintis are right in asserting that communities work well relative to markets and states where the tasks are qualitative and hard to capture in explicit contracts, and the conflicts of interest among the members are limited, it seems likely that extremely unequal societies will be competitively disadvantaged in the future because their structures of privilege and material reward limit the capacity of community governance to facilitate the qualitative interactions that underpin the modern economy. Political democracy, policies that limit the extent of social and economic inequality, and widespread civil liberties may thus not only be desirable in terms of political ethics, but may in fact be necessary to harness moral sentiments to future economic and social development around the world.

Notes

1. We say an action is *altruistic* when it confers benefits to other members of a group at a cost to the actor. Note that this definition says nothing about the intentions of the actor. Note also that an action can be altruistic yet increase the subjective utility of the actor. Indeed, any voluntary, intended act of altruism will have this property.

2. Since we care about behavior rather than its subjective correlates, throughout this chapter we use the term "self-regarding" rather than "self-interested." For instance, if one truly cares about others, it may be self-interested to sacrifice on their behalf, even though it is manifestly non–self-regarding to do so.

3. While the term "strong reciprocity" is new, the idea certainly is not, having been studied by Homans (1958), Gouldner (1960), Moore Jr. (1978), Frank (1988), and Hirshleifer and Rasmusen (1989), among others.

4. The adaptive significance of the human ability to detect cheaters was stressed by Cosmides and Tooby (1992) who, in contrast with our usage, consider this capacity as individually fitness-enhancing rather than altruistic. The precommitment to punish transgressors has been insightfully analyzed by Hirshleifer (1987) and Frank (1988).

5. Classical group selection involves the altruistic behavior having fitness costs as compared with behavior of non-altruistic group members, but these costs being more than offset by the higher fitness of groups with many altruists, as compared with groups in which altruism is rare or absent.

6. By *multilevel selection* (Keller 1999), we mean that selection operates at some level other than that of the gene or individual. For instance, the social organization of a beehive contributes to the fitness of individual bees, which leads to the growth of beehives.

7. The observed behavior was predicted by Akerlof (1982).

8. For additional experimental results and analysis, see Bowles and Gintis (2002) and Fehr and Gächter (2002).

References

Akerlof, George A. "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics* 97, 4 (November 1982): 543–569.

Alcock, John. *Animal Behavior: An Evolutionary Approach*. Sunderland, MA: Sinauer, 1993.

Alexander, Richard D. *The Biology of Moral Systems*. New York: Aldine, 1987.

Andreoni, James. "Cooperation in Public Goods Experiments: Kindness or Confusion," *American Economic Review* 85, 4 (1995): 891–904.

Andreoni, James, and John H. Miller. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica* 70, 2 (2002): 737–753.

Andreoni, James, Brian Erard, and Jonathan Feinstein. "Tax Compliance," *Journal of Economic Literature* 36, 2 (June 1998): 818–860.

Axelrod, Robert, and William D. Hamilton. "The Evolution of Cooperation," *Science* 211 (1981): 1390–1396.

Blount, Sally. "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences," *Organizational Behavior & Human Decision Processes* 63, 2 (August 1995): 131–144.

Boehm, Christopher. "The Evolutionary Development of Morality as an Effect of Dominance Behavior and Conflict Interference," *Journal of Social and Biological Structures* 5 (1982): 413–421.

Boone, James L., and Karen L. Kessler. "More Status or More Children? Social Status, Fertility Reduction, and Long-Term Fitness," *Evolution & Human Behavior* 20, 4 (July 1999): 257–277.

Boorman, Scott A., and Paul Levitt. *The Genetics of Altruism*. New York: Academic Press, 1980.

Bowles, Samuel. "Individual Interactions, Group Conflicts, and the Evolution of Preferences," in Steven N. Durlauf and H. Peyton Young (eds.) *Social Dynamics*. Cambridge, MA: MIT Press, 2001, 155–190.

Bowles, Samuel, and Herbert Gintis. "Homo Reciprocans," *Nature* 415 (10 January 2002): 125–128.

———. "The Evolution of Strong Reciprocity: Cooperation in Heterogeneous Populations," *Theoretical Population Biology* 65 (2004): 17–28.

Bowles, Samuel, Jung-kyoo Choi, and Astrid Hopfensitz. "The Co-evolution of Individual Behaviors and Social Institutions," *Journal of Theoretical Biology* 223 (2003): 135–147.

Boyd, Robert, and Peter J. Richerson. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press, 1985.

———. "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizeable Groups," *Ethology and Sociobiology* 113 (1992): 171–195.

———. "Group Beneficial Norms Can Spread Rapidly in a Cultural Population," *Journal of Theoretical Biology* 215 (2002): 287–296.

Bowles, Samuel, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. "Evolution of Altruistic Punishment," *Proceedings of the National Academy of Sciences* 100, 6 (March 2003): 3531–3535.

Brown, Donald E. *Human Universals*. New York: McGraw-Hill, 1991.

Camerer, Colin, and Richard Thaler. "Ultimatums, Dictators, and Manners," *Journal of Economic Perspectives* 9, 2 (1995): 209–219.

Coase, Ronald H. "The Problem of Social Cost," *Journal of Law and Economics* 3 (October 1960): 1–44.

Coleman, James S. *Foundations of Social Theory*. Cambridge, MA: Belknap, 1990.

Cosmides, Leda, and John Tooby. "Cognitive Adaptations for Social Exchange," in Jerome H. Barkow, Leda Cosmides, and John Tooby (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York: Oxford University Press, 1992, 163–228.

Crow, James F., and Motoo Kimura. *An Introduction to Population Genetic Theory*. New York: Harper & Row, 1970.

Dawes, Robyn M., and Richard Thaler. "Cooperation," *Journal of Economic Perspectives* 2 (1988): 187–197.

Dawkins, Richard. *The Selfish Gene*. Oxford: Oxford University Press, 1976.

———. *The Selfish Gene, 2nd Edition*. Oxford: Oxford University Press, 1989.

Durham, William H. *Coevolution: Genes, Culture, and Human Diversity*. Stanford: Stanford University Press, 1991.

Edgerton, Robert B. *Sick Societies: Challenging the Myth of Primitive Harmony*. New York: The Free Press, 1992.

Eibl-Eibesfeldt, I. "Warfare, Man's Indoctrinability and Group Selection," *Journal of Comparative Ethnology* 60, 3 (1982): 177–198.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. "Testing Theories of Fairness and Reciprocity-Intentions Matter," 2002. University of Zürich.

Fehr, Ernst, and Bettina Rockenbach. "Detrimental Effects of Incentives on Human Altruism?" *Nature* 422 (March 2003): 137–140.

Fehr, Ernst, and J. List. "The Hidden Costs and Returns of Incentives: Trust and Trustworthiness among CEOs," 2002. Working Paper, Institute for Empirical Research, University of Zürich.

Fehr, Ernst, and Simon Gächter. "Cooperation and Punishment," *American Economic Review* 90, 4 (September 2000a): 980–994.

———. "Do Incentive Contracts Crowd Out Voluntary Cooperation?" 2000b. Working Paper No. 34, Institute for Empirical Research, University of Zürich.

———. "Altruistic Punishment in Humans," *Nature* 415 (10 January 2002): 137–140.

Fehr, Ernst, and Klaus M. Schmidt. "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics* 114 (August 1999): 817–868.

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger. "Reciprocity as a Contract Enforcement Device: Experimental Evidence," *Econometrica* 65, 4 (July 1997): 833–860.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl. "Does Fairness Prevent Market Clearing?" *Quarterly Journal of Economics* 108, 2 (1993): 437–459.

———. "Gift Exchange and Reciprocity in Competitive Experimental Markets," *European Economic Review* 42, 1 (1998): 1–34.

Feldman, Marcus W., and Lev A. Zhivotovsky. "Gene-Culture Coevolution: Toward a General Theory of Vertical Transmission," *Proceedings of the National Academy of Sciences* 89 (December 1992): 11935–11938.

Feldman, Marcus W., Luca L. Cavalli-Sforza, and Joel R. Peck. "Gene-Culture Coevolution: Models for the Evolution of Altruism with Cultural Transmission," *Proceedings of the National Academy of Sciences* 82 (1985): 5814–5818.

Fix, Alan. *Migration and Colonization in Human Microevolution*. Cambridge: Cambridge University Press, 1999.

Foster, George M. *Culture and Conquest: America's Spanish Heritage*. New York: Wenner-Gren, 1960.

Frank, Robert H. *Passions Within Reason: The Strategic Role of the Emotions*. New York: Norton, 1988.

Frey, Bruno. "A Constitution for Knaves Crowds Out Civic Virtue," *Economic Journal* 107, 443 (July 1997): 1043–1053.

———. "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding Out," *American Economic Review* 87, 4 (September 1997): 746–755.

———. *Not Just for the Money: An Economic Theory of Personal Motivation*. Cheltenham, UK: Edward Elgar, 1997.

Friedman, Milton. *Essays in Positive Economics*. Chicago: University of Chicago Press, 1953.

Frohlich, Norman, and Joe Oppenheimer. "Choosing Justice in Experimental Democracies with Production," *American Political Science Review* 84, 2 (June 1990): 461–477.

Fudenberg, Drew, and Eric Maskin. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica* 54, 3 (May 1986): 533–554.

Gächter, Simon, and Ernst Fehr. "Collective Action as a Social Exchange," *Journal of Economic Behavior and Organization* 39, 4 (July 1999): 341–369.

Gintis, Herbert. "Strong Reciprocity and Human Sociality," *Journal of Theoretical Biology* 206 (2000): 169–179.

———. "Solving the Puzzle of Human Prosociality," *Rationality and Society* 15, 2 (May 2003): 155–187.

———. "The Hitchhiker's Guide to Altruism: Genes, Culture, and the Internalization of Norms," *Journal of Theoretical Biology* 220, 4 (2003): 407–418.

Gintis, Herbert, Eric Alden Smith, and Samuel Bowles. "Costly Signaling and Cooperation," *Journal of Theoretical Biology* 213 (2001): 103–119.

Glaeser, Edward, David Laibson, Jose A. Scheinkman, and Christine L. Soutter. "Measuring Trust," *Quarterly Journal of Economics* 65 (2000): 622–846.

Gouldner, Alvin W. "The Norm of Reciprocity: A Preliminary Statement," *American Sociological Review* 25 (1960): 161–178.

Greenberg, M. S., and D. M. Frisch. "Effect of Intentionality on Willingness to Reciprocate a Favor," *Journal of Experimental Social Psychology* 8 (1972): 99–111.

Güth, Werner, and Reinhard Tietz. "Ultimatum Bargaining Behavior: A Survey and Comparison of Experimental Results," *Journal of Economic Psychology* 11 (1990): 417–449.

Hamilton, W. D. "The Genetical Evolution of Social Behavior," *Journal of Theoretical Biology* 37 (1964): 1–16, 17–52.

Hechter, Michael, and Satoshi Kanazawa. "Sociological Rational Choice," *Annual Review of Sociology* 23 (1997): 199–214.

Henrich, Joseph, and Robert Boyd. "Why People Punish Defectors: Weak Conformist Transmission Can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas," *Journal of Theoretical Biology* 208 (2001): 79–89.

Henrich, Joe, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. *Foundations of Human Sociality: Ethnography and Experiments in Fifteen Small-scale Societies*. Oxford: Oxford University Press, 2004.

Henrich, Joe, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. "Cooperation, Reciprocity and Punishment in Fifteen Small-scale Societies," *American Economic Review* 91 (May 2001): 73–78.

Hertwig, Ralph, and Andreas Ortmann. "Experimental Practices in Economics: A Methodological Challenge for Psychologists?" *Behavioral and Brain Sciences* 24 (2001): 383–451.

Hirschman, Albert. "Against Parsimony," *Economic Philosophy* 1 (1985): 7–21.

Hirshleifer, David, and Eric Rasmusen. "Cooperation in a Repeated Prisoners' Dilemma with Ostracism," *Journal of Economic Behavior and Organization* 12 (1989): 87–106.

Hirshleifer, Jack. "Economics from a Biological Viewpoint," in Jay B. Barney and William G. Ouchi (eds.) *Organizational Economics*. San Francisco: Jossey-Bass, 1987, 319–371.

Homans, George C. "Social Behavior as Exchange," *American Journal of Sociology* 65, 6 (May 1958): 597–606.

Hume, David. *Essays: Moral, Political and Literary*. London: Longmans, Green, 1898(1754).

Keckler, C. N. W. "Catastrophic Mortality in Simulations of Forager Age-of-Death: Where Did all the Humans Go?" in R. Paine (ed.) *Integrating Archaeological Demography: Multidisciplinary Approaches to Prehistoric Populations*. Center for Archaeological Investigations, Occasional Papers No. 24, 205–228.

Keller, Laurent. *Levels of Selection in Evolution*. Princeton, NJ: Princeton University Press, 1999.

Kelly, Raymond C. *The Nuer Conquest: The Structure and Development of an Expansionist System*. Ann Arbor: University of Michigan Press, 1985.

Knauft, Bruce. "South Coast New Guinea Cultures: History, Comparison, Dialectic," *Cambridge Studies in Social and Cultural Anthropology* 89 (1993).

Ledyard, J. O. "Public Goods: A Survey of Experimental Research," in J. H. Kagel and A. E. Roth (eds.) *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press, 1995, 111–194.

Loewenstein, George. "Experimental Economics from the Vantage Point of View of Behavioural Economics," *Economic Journal* 109, 453 (February 1999): F25–F34.

Lumsden, C. J., and E. O. Wilson. *Genes, Mind, and Culture: The Coevolutionary Process*. Cambridge, MA: Harvard University Press, 1981.

Maynard Smith, John. "Group Selection," *Quarterly Review of Biology* 51 (1976): 277–283.

Monroe, Kristen Renwick. *The Economic Approach to Politics*. Reading, MA: Addison Wesley, 1991.

Moore, Jr., Barrington. *Injustice: The Social Bases of Obedience and Revolt*. White Plains: M. E. Sharpe, 1978.

Orbell, John M., Robyn M. Dawes, and J. C. Van de Kragt. "Organizing Groups for Collective Action," *American Political Science Review* 80 (December 1986): 1171–1185.

Ostrom, Elinor, James Walker, and Roy Gardner. "Covenants with and without a Sword: Self-Governance Is Possible," *American Political Science Review* 86, 2 (June 1992): 404–417.

Posner, Richard. *Economic Analysis of Law*. New York: Little, Brown, 1973.

Price, G. R. "Selection and Covariance," *Nature* 227 (1970): 520–521.

———. "Extension of Covariance Selection Mathematics," *Annals of Human Genetics* 35 (1972): 485–490.

Richerson, Peter J., and Robert Boyd. *The Nature of Cultures*. Chicago: University of Chicago Press, 2003.

Roof, Wade Clark, and William McKinney. *American Mainline Religion: Its Changing Shape and Future*. New Brunswick, NJ: Rutgers University Press, 1987.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review* 81, 5 (December 1991): 1068–1095.

Sato, Kaori. "Distribution and the Cost of Maintaining Common Property Resources," *Journal of Experimental Social Psychology* 23 (January 1987): 19–31.

Smith, Adam. *The Theory of Moral Sentiments*. Indianapolis: Liberty Fund, 1982(1759).

Smith, Adam. *The Wealth of Nations*. New York: Prometheus Books, 1991(1776).

Sober, Elliot, and David Sloan Wilson. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press, 1998.

Soltis, Joseph, Robert Boyd, and Peter J. Richerson. "Can Group-functional Behaviors Evolve by Cultural Group Selection: An Empirical Test," *Current Anthropology* 36, 3 (June 1995): 473–483.

Stephens, W., C. M. McLinn, and J. R. Stevens. "Discounting and Reciprocity in an Iterated Prisoner's Dilemma," *Science* 298 (13 December 2002): 2216–2218.

Taylor, Michael. *Anarchy and Cooperation*. London: John Wiley and Sons, 1976.

Titmuss, R. M. *The Give Relationship*. London: Allen and Unwin, 1970.

Trivers, R. L. "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology* 46 (1971): 35–57.

Williams, G. C. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton, NJ: Princeton University Press, 1966.

Wilson, Edward O. *Sociobiology: The New Synthesis*. Cambridge, MA: Harvard University Press, 1975.

Yamagishi, Toshio. "The Provision of a Sanctioning System as a Public Good," *Journal of Personality and Social Psychology* 51 (1986): 110–116.

———. "The Provision of a Sanctioning System in the United States and Japan," *Social Psychology Quarterly* 51, 3 (1988a): 265–271.

———. "Seriousness of Social Dilemmas and the Provision of a Sanctioning System," *Social Psychology Quarterly* 51, 1 (1988b): 32–42.

———. "Group Size and the Provision of a Sanctioning System in a Social Dilemma," in W. B. G. Liebrand, David M. Messick, and H. A. M. Wilke (eds.) *Social Dilemmas: Theoretical Issues and Research Findings*. Oxford: Pergamon Press, 1992, 267–287.